Instituto Federal de Educação, Ciência e Tecnologia da Paraíba

Campus João Pessoa

Programa de Pós-Graduação em Tecnologia da Informação

Nível Mestrado Profissional

JEFFERSON DE MORAIS TOLEDO

PREDICTING OCCUPATIONAL ACCIDENTS IN BRAZIL: A MACHINE LEARNING BASED APPROACH

DISSERTAÇÃO DE MESTRADO

JOÃO PESSOA

2025

Jefferson de Morais Toledo

Predicting occupational accidents in Brazil: a machine learning based approach

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós-Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB.

Orientador: Prof. Dr. Thiago José Marques Moura

João Pessoa

2025

Dados Internacionais de Catalogação na Publicação (CIP)
Biblioteca Nilo Peçanha - *Campus* João Pessoa, PB.

T649p Toledo, Jefferson de Morais.

Predicting occupational accidents in Brazil: a machine learning based approach / Jefferson de Morais Toledo. – 2025.
71 f. : il.

Dissertação (Mestrado em Tecnologia da Informação) – Instituto Federal de Educação da Paraíba / Programa de Pós-Graduação em Tecnologia da Informação (PPGTI), 2025.
Orientação: Profº. Dr. Thiago José Marques Moura.

1.Acidente de trabalho. 2. Aprendizado de máquina. 3. Algoritmos de regressão. I. Título.

CDU 331.46:004.8(043)

Bibliotecária responsável Lucrecia Camilo de Lima – CRB 15/132

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA

PROGRAMA DE PÓS-GRADUAÇÃO *STRICTO SENSU*

MESTRADO PROFISSIONAL EM TECNOLOGIA DA INFORMAÇÃO

JEFFERSON DE MORAIS TOLEDO

PREDICTING OCCUPATIONAL ACCIDENTS IN BRAZIL: A MACHINE LEARNING BASED APPROACH

Dissertação apresentada como requisito para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós- Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB - Campus João Pessoa.

Aprovado em 06 de junho de 2025

Membros da Banca Examinadora:

Dr. Thiago José Marques Moura

IFPB - PPGTI

Dra. Damires Yluska Souza Fernandes

IFPB - PPGTI

Dr. Yuri de Almeida Malheiros Barbosa

UFPB

João Pessoa/2025

Ao meu pai.

RESUMO

Acidentes de trabalho representam um sério problema social no Brasil, muitas vezes causando mortes de trabalhadores ou incapacidade permanente para o trabalho. Portanto, é necessário estudar a distribuição estatística desses acidentes no país e avaliar a possibilidade de prever a ocorrência desses fenômenos. Neste trabalho, obtemos um conjunto de dados unificado, que chamamos de BrStats, com dados estatísticos de todas as cidades brasileiras, integrando dados públicos relacionados à população, economia, educação e saúde. Em seguida, usamos o BrStats e adicionamos mais dados relacionados a acidentes de trabalho, inspeções trabalhistas e emprego para obter conjuntos de dados que são usados para treinar modelos de aprendizado de máquina (ML) para prever a ocorrência de acidentes de trabalho no Brasil. Primeiramente, prevemos o número de acidentes de trabalho em cada atividade econômica nos estados brasileiros treinando modelos de regressão - regressão linear, máquina de vetores de suporte (SVM), XGBoost e LightGBM. Neste cenário, obtemos valores de R^2 próximos a 0,9. Prevemos, então, o número de acidentes de trabalho em cidades brasileiras usando regressão linear como baseline, além de algoritmos de árvores de decisão e gradient boosting (GradientBoosting, LightGBM, XGBoost e CatBoost), com a adoção de otimização bayesiana para ajustar os hiperparâmetros dos algoritmos. Também usamos uma técnica de seleção de features baseada na importância das features para avaliar quais variáveis apresentam maior significância nas predições. Os modelos apresentam alto desempenho, com a métrica R-quadrado atingindo valores superiores a 0,90. Os resultados obtidos neste trabalho podem auxiliar o governo e as empresas a adotarem ações preventivas para evitar acidentes de trabalho, reduzindo os custos humanos e previdenciários do país.

Palavras-chaves: Acidentes de trabalho. Aprendizado de máquina. Algoritmos de regressão.

ABSTRACT

Occupational accidents represent a serious social problem in Brazil, often causing deaths of workers or permanent incapacity to work. Therefore, it is necessary to study the statistical distribution of such accidents in the country and evaluate the possibility of predicting their occurrence. In this work, we obtain a unified dataset, which we call BrStats, with statistical data for all Brazilian cities, integrating public data related to population, economy, education, and health. Then, we use BrStats and add more data related to occupational accidents, labor inspections, and employment to obtain datasets that are used to train machine learning (ML) models to predict the occurrence of occupational accidents in Brazil. Firstly, we predict the number of occupational accidents in each economic activity in Brazilian states by training regression models - linear regression, support vector machine (SVM), XGBoost, and LightGBM. In this scenario, we obtain R^2 values near 0.9. We, then, predict the number of occupational accidents in Brazilian cities using linear regression as a baseline predictor in addition to Decision Tree (DT) and gradient boosting algorithms (GradientBoosting, LightGBM, XGBoost, and CatBoost), with the adoption of Bayesian optimization to tune the algorithms' hyperparameters. We also use a feature selection technique based on the importance of the variables to evaluate which features present the greatest significance in the predictions. The models show high performance, with the R-squared metric reaching values greater than 0.90. The results obtained in this work may help the government and enterprises adopt preventive actions to avoid occupational accidents, reducing the country's human and social security costs.

Key-words: Occupational accidents. Machine learning. Regression algorithms.

LIST OF FIGURES

LIST OF TABLES

LISTA DE ABREVIATURAS E SIGLAS

AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Programming Interface
BN	Bayesian Networks
CBO	Código Brasileiro de Ocupações
CID	Classificação Estatística Internacional de Doenças
CNAE	Cadastro Nacional de Atividades Econômicas
DT	Decision Trees
GDP	Gross Domestic Product
ETL	Extract, Transform, and Load
HDI	Human Development Index
IBGE	Instituto Brasileiro de Geografia e Estatística
IDEB	Índice de Desenvolvimento da Educação Básica
ILO	International Labor Organization
IPEA	Instituto de Pesquisa Econômica Aplicada
KNN	K-nearest neighbors
LR	Logistic Regression
ML	Machine Learning
MLP	Multilayer Perceptron
NB	Naive Bayes
RF	Random Forest
RMSE	Root Mean Squared Error
SVC	Support Vector Classifier
SVM	Support Vector Machines

SVR	Support Vector Regressor
WHO	World Health Organization

SUMÁRIO

1 INTRODUCTION

In this chapter, we provide a brief introduction to the work, defining the motivation and the proposed problem. We also list the objectives of the research developed and list the published works.

1.1 Motivation and problem statement

According to Brazilian legislation, an occupational accident occurs while an employee works for a company, causing injury or functional disturbance that leads to death or the permanent or temporary loss or reduction of the capacity to work. In one decade (from 2013 to 2022), 6,090,076 (six million, ninety thousand, and seventy-six) occupational accidents and work-related diseases were reported in Brazil. In the same period, 23,649 (twenty-three thousand, six hundred and forty-nine) workers lost their lives due to work-related causes (MPT, 2023). In 2021, the social security expenses estimated to result from occupational accidents and diseases exceeded 102 billion reais (MPT, 2023) (approximately 18,7 billion dollars in 2024 exchange rate), a significant portion of the Brazilian Gross Domestic Product (GDP).

The International Labor Organization (ILO) and the World Health Organization (WHO) estimate that occupational accidents and work-related diseases cause the death of almost two million workers per year (ORGANIZATION et al., 2021) in the world. In 2016, occupational accidents and diseases caused the death of 1.9 million people, overloading the countries' health systems, reducing family income, and decreasing economic productivity (ORGANIZATION et al., 2021). It is important to mention that the main reason for deaths at work was exposure to long working hours, which caused the death of 750,000 employees annually (ORGANIZATION et al., 2021).

The large number of occupational accidents represents a serious public health problem worldwide, especially in Brazil. These accidents can, however, be prevented (IVASCU; CIOCA, 2019; ALLI, 2008) and, according to Brazilian law, the government is responsible for monitoring labor laws and ensuring safe environments for workers. Government agencies can apply many methods, including technology, to achieve these objectives.

Recently, we have observed growth in the use of data in planning and decision-making processes both in the public (SETHI et al., 2025; MERGEL; RETHEMEYER; ISETT, 2016; MACIEJEWSKI, 2017) and in the private sectors (WANG; ZONG, 2023; LAAT, 2018). The increase in storage and the power of computer processing are allowing improvement in data mining. The consequent availability of data is bringing benefits to society, increasing the accuracy of the decision-making process, accelerating the planning stages in companies, and reducing costs (MACIEJEWSKI, 2017).

Machine learning has become a common expression since computers' increasing data availability and processing powers are stimulating the development of predictive models (ALPAYDIN, 2021). Many applications use ML algorithms, which are also used to solve problems in many fields of knowledge (ALPAYDIN, 2021). Thus, the government can apply this technology to implement public policies for health and safety at work.

In this work, we aim to deal with some key research questions. 1) How do we extract sociodemographic variables of Brazilian cities from multiple sources? 2) What is the statistical behavior of occupational accidents in Brazil? 3) Which of these variables can describe the occurrence of occupational accidents in the country? 4) Which ML model best explains the number of accidents through these variables? This work is developed as a means to answer these questions.

In this sense, one further problem still needs to be addressed: which variables can be used as features to predict occupational accidents in the country, and from which sources can we obtain these data? Brazil has 5,570 municipalities (IBGE, 2023a) spread in an area of more than 8 million square kilometers (IBGE, 2023d). In this vast territory, the country shows a diversity of climates (from temperate to equatorial), biomes (from the Amazon rain forest to semi-arid), and social behaviors (IBGE, 2023c; IBGE, 2023b). While Brazil's largest city (São Paulo) has more than 12 million inhabitants, some municipalities have as few as a thousand people. Given these characteristics, obtaining socioeconomic variables of the country's cities is not a simple task. As an initial step, we propose extracting data from multiple sources to get a unified dataset with socioeconomic variables for Brazilian cities.

In this work, we also propose using machine-learning regression algorithms to predict the number of occupational accidents in each economic activity in Brazilian states and to foresee the number of occupational accidents in all Brazilian cities. The main goal of this study is to improve preventive measures for Brazil's public and private sectors, promote a safe workplace, and avoid occupational accidents.

1.2 Objectives

Now, let us identify the research's general and specific objectives.

1.2.1 General objective

The general objective of this work is to predict the number of occupational accidents in Brazil using machine learning algorithms and socioeconomic variables as features.

1.2.2 Specific objectives

Our initial objective in developing the research is to obtain a unified dataset with socioeconomic variables for all Brazilian cities. This work also analyzes which data sources can be used

to obtain statistical data for all Brazilian cities. Then, we obtain a unified dataset with population, economy, employment, education, and health variables, performing an extensive exploratory analysis of the occupational accident data and the socioeconomic variables used as features. We also intend to use machine learning models to predict the occurrence of occupational accidents in each Brazilian state, using the cited socioeconomic variables as features. Finally, we predict the number of occupational accidents in each Brazilian cities, evaluating which features are the most important in the models' decisions and comparing the models' predictions, performing a statistical hypothesis test.

1.3 Document structure

This document is organized as follows. In Chap. 2, we summarize the works related to the object of study of this research and briefly review some fundamental concepts about machine learning. Chap. 3 is devoted to obtaining a unified dataset, which we call BrStats, from multiple open-source datasets. This dataset contains statistical variables that are used as features in subsequent chapters. In Chap. 4, we propose using ML regression models to predict the number of occupational accidents in each economic activity of Brazilian states. We predict the occurrence of occupational accidents in Brazilian cities in Chap. 5. Finally, Chap. 6 presents the concluding remarks and possible future works. It is important to note that Chaps. 3, 4, and 4 are based on papers written during this research.

1.4 Articles written during the research

Some articles were presented at international conferences and published or submitted to academic journals during this research. Tab. 1.

Reference	**Title**	**Description**	**Conference/ journal**
(TOLEDO; MOURA; TIMOTEO, 2023)	BrStats: a socioeconomic statistics dataset of the Brazilian cities	We obtain a unified dataset with statistical data for all cities in the country, integrating data related to population, economy, employment, education, and health	SBBD-DSW
(TOLEDO; MOURA, 2024)	Occupational Accidents Prediction in Brazilian States: A Machine Learning Based Approach	We predict the number of occupational accidents in each economic activity in Brazilian states	ICEIS 2025 (A3)
(TOLEDO; MOURA, 2025a)	Occupational accidents prediction in Brazil: an approach using regression Machine learning algorithms	We predict the number of occupational accidents in Brazilian states (extended version)	Lecture Notes in Business Information Processing (B1)
(TOLEDO; MOURA, 2025b)	Predicting occupational accidents in Brazilian cities: a machine learning based approach	We predict the number of occupational accidents in Brazilian cities (submitted paper)	International Journal of Data Science and Analytics (A1)

Table 1 – Papers produced during the research period.

2 THEORETICAL FOUNDATION

In this chapter, we present the theoretical foundations of the developed research and the related papers that supported this work.

2.1 Machine learning

In recent years, the increase in computational capacity and data storage has driven the development of machine learning, a field of study that is part of artificial intelligence (AI), representing the intersection of statistics, mathematics, and computing (MITCHELL, 1997). Government entities and private organizations have used this technology to plan and provide their products and services.

Arthur Samuel defined machine learning as the "field of study that gives computers the ability to learn without being explicitly programmed" (SAMUEL, 1959). As Alpaydin (2021) stated, "With machine learning, data begins to guide operations; it is no longer the programmers, but the data that defines what to do next" (our translation). Machine learning, therefore, is related to the ability of computers to learn from previous data and make predictions that improve as the systems are fed with new data (MITCHELL, 1997). In recent years, the increase in computing and data storage capacity has driven the development of machine learning, an area of knowledge integral to artificial intelligence (AI) that represents the intersection of statistics, mathematics, and computing. Several government entities have used the technology in their activities, and private companies have applied ML algorithms to provide their products and services. Examples of solutions developed with ML include image recognition and spam detection applications.

Continuing on the topic, it is worth highlighting a publication in the prestigious journal Science in which Jordan e Mitchell (2015) state that machine learning "is expected to be one of the most transformative technologies of the 21st century" (our translation) and that, therefore, it deserves to be widely studied and implemented.

2.1.1 Supervised and unsupervised learning

Most computational learning problems can be classified into two categories: supervised learning problems and unsupervised learning problems (JAMES et al., 2013). In supervised learning, we intend to predict a variable (known as a dependent variable, output, or target) based on one or more input variables (also called independent variables or features). In unsupervised learning, there are no target variables to estimate and the algorithms perform pattern detection in the independent variables with the aim, for example, of performing automated segmentation of data into classes (clusters), reducing the dimensionality of the training data set, or detecting atypical values (outliers) in this set.

Supervised learning problems can be divided into regression problems and classification problems. While in regression, the target variable assumes continuous numerical values, in classification problems, the target variable is contained in a finite and discrete set of values.

In general, supervised learning is a mathematical method by which it is possible to, given a training dataset $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$, find a function $f(\mathbf{x})$, with $\mathbf{x} = (x_1, x_2, ..., x_N)$, that maps the features, $\mathbf{x}$ into a target variable $y = (y_1, y_2, ..., y_N)$ (JAMES et al., 2013; ALPAYDIN, 2021). The predicted values of the target variable, $\hat{y}$, are determined by the ML algorithm in such a way that it minimizes an objective function given by

$$obj = L(\theta) + \Omega(\theta), \tag{1}$$

where θ represents the model parameters, $L(\theta)$ is the training loss, which measures how well the model performs the predictions. $\Omega(\theta)$ is the regularization term, which is added to avoid overfitting and is related to the complexity of the model.

In this work, as we discuss in what follows, we intend to predict the number of occupational accidents and economic activity in each Brazilian state and city. Therefore, the problem's target variable is a continuous number, and we must use regression algorithms. Thus, let us briefly describe the ML regression algorithms implemented in the proposed experimental protocol.

2.1.2 Machine learning algorithms

Now, let us briefly describe the ML models used in this work. As already mentioned, we use regression models in our research, so we focus on this kind of algorithm in this subsection. The algorithms used in this work were chosen based on the related works described below and are widely used in the literature and ML competitions.

2.1.2.1 Linear regression

The simplest regression model is called linear regression (JAMES et al., 2013). It assumes approximately a linear relationship between the features, $\mathbf{x} = (x_1, x_2, ..., x_N)$, and the target variable, $y = (y_1, y_2, ..., y_N)$, such that the predicted variable $\hat{y}$ can be obtained by

$$\hat{y} = \beta_0 + \beta_1 \mathbf{x}, \tag{2}$$

where β_0 and β_1 are coefficients that are estimated using data and that minimize the discrepancies between predicted and actual output values. These coefficients can be calculated by

$$\beta_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}, \tag{3}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}, \tag{4}$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_1$ are the mean values of $\mathbf{x}$ and y.

Although simple compared to more modern models described in the following, this model is widely used in science (JAMES et al., 2013). We use linear regression as a baseline model, a starting point for the more complex algorithms described below.

2.1.2.2 Support Vector Algorithms

Support Vector Machines are a popular class of ML models developed in the 1990s and introduced for classification problems (JAMES et al., 2013). The SVM models were generalized to other types of problems and are currently used in various application domains, from text categorization to computer vision (MAMMONE; TURCHI; CRISTIANINI, 2009).

In simplified terms, the original support vector algorithms seek to find an optimized hyperplane that separates two classes in a dataset, maximizing the margin, i.e., the minimal distance between the hyperplane and the closest data points (MAMMONE; TURCHI; CRISTIANINI, 2009). The introduction of kernel functions in the algorithms makes it possible to draw nonlinear decision surfaces between the classes, generalizing the method (MAMMONE; TURCHI; CRISTIANINI, 2009).

2.1.2.3 Decision trees

Decision trees (DT) are tree-based algorithms used in both classification or regression problems (JAMES et al., 2013). In summary, a regression tree recursive split training data in such a way that the samples with similar target values are grouped (JAMES et al., 2013). The algorithm searches for the optimal split points within a tree. In this work, we use the Classification and Regression Tree (CART) algorithm in the DT implementation (BREIMAN, 2017). We use DT as a simpler tree-based model to compare with the gradient boosting algorithms explained below.

2.1.2.4 Gradient boosting algorithms

Boosting algorithms use an iterative sequence of weak learners, i.e., slightly better than random guessing, to obtain a strong learner (FREUND; SCHAPIRE; ABE, 1999; SCHAPIRE et al., 1999). If the base learners are decision trees, these models are called boosted trees or tree boosting.

Gradient boosting trees are algorithms obtained by interactively adding the predictions of each tree. If we consider that the prediction of each tree is given by $f_i(\mathbf{x})$, in a boosted tree algorithm with K trees, the predicted output is given by (CHEN; GUESTRIN, 2016; BROWNLEE, 2019)

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) = \hat{y}_i^{K-1} + f_K(x_i). \tag{5}$$

If the base learners are trees, we can write each f_i as (CHEN; GUESTRIN, 2016):

$$f_i(\mathbf{x}) = \mathbf{w}_{q(\mathbf{x})}, \quad w \in R^T, \quad q : R^d \to \{1, 2, ..., T\}, \tag{6}$$

where q is a function that maps the features into one leaf and $\mathbf{w}$ is a vector of scores on leaves, T is the number of leaves, and d is the dimension of each data point. In this scenario, we can also write the objective function, Eq. (1), as

$$obj = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{7}$$

Instead of directly solving the optimization problem of reducing the loss in Eq. (7), the parameter adjustment in Eq. (6) is done by the gradient descent algorithm (BENTÉJAC; CSÖRGŐ; MARTÍNEZ-MUÑOZ, 2021).

XGBoost Extreme Gradient Boosting (XGBoost) is another gradient boosting algorithm developed focused on efficiency and flexibility, providing parallel tree boosting that solves ML problems in a fast and accurate way (CHEN; GUESTRIN, 2016).

In this algorithm, the regularization term of the objective function of Eq. (7) is given by

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2. \tag{8}$$

The regularization is an important aspect of XGBoost to prevent overfitting and also to reduce the time for training the model, given the reduction of the complexity of the algorithm (BENTÉJAC; CSÖRGŐ; MARTÍNEZ-MUÑOZ, 2021).

XGBoost is also recognized by winning ML competitions (NIELSEN, 2016) and building trees in a parallel way, using the CPU cores of the computers, contrasting with the traditional Gradient Boosting algorithms (NOBRE; NEVES, 2019).

LightGBM LightGBM is a gradient boosting tree algorithm developed by Microsoft, focusing on efficiency and scalability (KE et al., 2017). Compared to other boosting trees, LightGBM saves time and computational cost, allowing researchers and developers to deal with big datasets (SCHAPIRE et al., 1999).

In general, gradient boosting algorithms need to scan all data for each feature, estimating the information gain of all possible split points (KE et al., 2017). This process is very time-consuming for large datasets. To avoid scanning all the instances of the training dataset, LighGBM

introduced new methods: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) (KE et al., 2017).

The feature space **x** is, mathematically, a vector space. It can be verified that instances of **x** with larger vector gradient contribute more to the information gain, reducing the objective function value of Eq. (7). The GOSS method down-samples the dataset instances, keeping those with large gradients, and only randomly dropping those instances with small gradients (KE et al., 2017).

In real large datasets, the feature space is quite sparse, i.e., many features are almost exclusive, rarely taking nonzero values simultaneously (KE et al., 2017). Thus, LightGBM uses the EFB algorithm, through which it safely bundles exclusive features using graph theory (KE et al., 2017).

The robust results presented by the algorithm, added to the reduced computational cost as explained, led us to use LightGBM in the research development.

Catboost We also use the CatBoost algorithm (DOROGUSH; ERSHOV; GULIN, 2018; PROKHORENKOVA et al., 2018) in this work. This model uses innovative ordered boosting and a new algorithm for processing categorical features (PROKHORENKOVA et al., 2018), outperforming other implementations of gradient boosting (DOROGUSH; ERSHOV; GULIN, 2018). CatBoost has also been compared with other ML algorithms and has shown better performance in various types of problems (IBRAHIM et al., 2020; BAIG et al., 2023).

Thus, in this work, we utilize the aforementioned state-of-the-art algorithms to predict the number of occupational accidents in Brazil, obtaining relevant results as described below.

2.1.3 Bayesian optimization for hyperparameter tuning

While training, ML algorithms search for parameters that reduce the prediction error in the training dataset (JAMES et al., 2013). The models' performance can also be increased by a class of variables called hyperparameters, which can be set while the design of the model is being developed (LIASHCHYNSKYI; LIASHCHYNSKYI, 2019; FEURER; HUTTER, 2019; WU et al., 2019).

Some traditional methods for searching optimal hyperparameters, such as grid search, can be computationally expensive (WU et al., 2019). Thus, seeking to save computational costs in solving problems that deal with finding the maximum/minimun of a function, and saving computational costs, Bayesian optimization is based on Bayes' theorem and searches for the optimal point of a function, based on its prior values (WU et al., 2019). The method has been tested for ML model hyperparameterizations (WU et al., 2019), yielding good results in terms of time cost. The technique has also shown good results in ML competitions compared with hyperparameter tuning approaches (TURNER et al., 2021). Thus, we use Bayesian optimization for hyperparameter tuning in this work.

2.1.4 Regression models evaluation metrics

Given the variety of available machine learning models, deciding which method produces the best results in a given dataset is an important task (JAMES et al., 2013). Thus, let us briefly summarize the metrics used in this work to evaluate the prediction of the trained regression models.

Root mean square error (RMSE). The RMSE can be calculated as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{9}$$

where N is the number of prediction, y_i are the actual values of the variable and $\hat{y}_i$ are the predict values. It represents the square root of the squared differences between the actual and predicted values of a variable. The closer the RMSE is to zero, the better the predictions.

Mean absolute percentage error (MAPE). The MAPE is given by

$$MAPE = \frac{1}{N}\left|\sum_{i=1}^{N}\frac{y_i - \hat{y}_i}{y_i}\right|, \tag{10}$$

and represents the mean percentual difference between the predicted and actual value of a variable.

R-squared (R^2) or coefficient of determination. The R^2 is the proportion of variance in the target that the features can explain and can be given by (JAMES et al., 2013)

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}, \tag{11}$$

where $\bar{y}$ is the already defined mean of the vector y.

From Eq. (11), we can observe that the values of R^2 ranges between 0 and 1. We can also state that R^2 quantifies the difference between the predictions and guessing all the values of the target variable by its means. So, the greater the value of this metric, the more explainable the target variable is by the features through the regression model.

2.2 Related works

In recent years, some works have been produced using machine learning techniques on themes related to workers' health, with data on occupational accidents and, to a lesser extent, work-related diseases.

In Brazil and several other countries, workers are required to undergo medical exams that certify their health conditions for work. In this sense, Charapaqui-Miranda et al. (2019)

obtained a dataset composed of information from admission examinations of workers in Peru. After preprocessing the data, excluding people for whom there was a lack of information on laboratory tests, the authors used logistic regression. They obtained a binary classification model for predicting work capacity with an accuracy greater than 68% (sixty-eight percent) and AUC (Area Under the Curve) of about 60% (sixty percent).

It is possible to verify that several studies have been conducted to determine the consequences of occupational accidents based on their characteristics. Sarkar et al. (2019), for example, predicted whether an accident caused damage to workers or property with an accuracy of around 90% (ninety percent) by performing tests with SVM (Support Vector Machines) and ANN (Artificial Neural Networks). The authors used GA (genetic algorithm) and PSO (particle swarm optimization) algorithms to refine the hyperparameters of the models (SARKAR et al., 2019). In turn, Recal e Demirel (2021) used logistic regression, SVM, ANN, and SGB (Stochastic Gradient Boosting) to classify work accidents that occurred in the construction industry in Turkey. They worked in two scenarios: binary prediction (fatal accident or not) and prediction in three classes (simple, severe, or fatal accident). The analysis of the authors' results allows us to verify that the SVM and SGB algorithms performed better in the two-class problem, while the SGB obtained better metrics in the three-class problem (RECAL; DEMIREL, 2021). In addition, the authors state that the predictions in the class of fatal accidents surpassed the results of other classes in accuracy, which reveals that the selected features have characteristics associated with the severity of accidents and, therefore, the trained models can be used to prevent future occurrences (RECAL; DEMIREL, 2021).

Khairuddin et al. (2022) analyzed a public OSHA (Occupational Safety and Health Administration) database with 66,405 (sixty-six thousand, four hundred and five accidents) using five machine learning algorithms: SVM (support vector machines), KNN (K-Nearest Neighbors), Naïve Bayes, Decision Tree, and Random Forest. The authors employed a feature optimization technique, which retains only the three most important features in the models' training process. Using the described methodology, the authors could predict the possibility of hospitalization with 89% accuracy (eighty-nine percent) and with 95% accuracy (ninety-five percent) the occurrence of amputation as a result of an accident at work (KHAIRUDDIN et al., 2022).

Machine learning models were also used for predictions in some specific economic activities. Koc, Ekmekcioğlu e Gurgun (2021), for example, used data from approximately 48,000 accidents in civil construction in Turkey and, after a data preprocessing process through the encoding of categorical variables and the normalization of numerical variables, the authors predicted the possibility of permanent disability of the injured workers with an accuracy of 82% (eighty-two percent) through the application of the algorithm XGBoost (Extreme Gradient Boosting) and with the use of a genetic algorithm to fix the hyperparameters of the model (KOC; EKMEKCIOĞLU; GURGUN, 2021).

In another work, Scott et al. (2021) used prehospital care data to predict which admissions

occurred due to occupational accidents in rural areas. Intending to help reduce the underreporting of occupational accidents, the authors used the Naïve Bayes algorithm and claimed to reduce by 69% (sixty-nine percent) the need for visual inspection of pre-hospital care cases (SCOTT et al., 2021). In the medical-hospital activity, Koklonis et al. (2021) used post-accident (or post-incident) data to classify events into five classes: needle/cut accident, fall, incident, accident, and safe condition. The authors categorized the data into the classes above with an accuracy of 93% (ninety-three percent), performing tests with the Naïve Bayes, MLP (multilayer perceptron), KNN, and BN (Bayesian Networks) algorithms (KOKLONIS et al., 2021).

In Brazil, the Labor Inspectors created a binary classification model for accidents that could create a probability of occurrence of accidents (TOLEDO; TIMOTEO; BARBOSA, 2020). The trained model presented an 86% (eighty-six percent) accuracy in the test dataset and the generated probabilities have been used in the planning of inspections in the country (TOLEDO; TIMOTEO; BARBOSA, 2020).

It is also possible to verify studies that use natural language processing techniques before the application of supervised learning algorithms (CHENG; KUSOEMO; GOSNO, 2020; GANGULI; MILLER; POTHINA, 2021). Based on an OSHA construction accident description database, Cheng, Kusoemo e Gosno (2020) built a model to classify texts into one of eleven classes (fall, impact against a moving object, explosion, etc.). The texts were preprocessed using the usual techniques of natural language processing (NLP): cleaning and removal of connectives, tokenization, and embedding. In the classification stage, several models were trained such as KNN, SVM, and neural networks, obtaining an accuracy of up to 0.71 (CHENG; KUSOEMO; GOSNO, 2020). Similarly, Ganguli et al. used data from the American Mine Safety and Health Administration (MSHA) containing the description of accidents that occurred in the mining sector to perform their classification between the types of overexertion when lifting objects, overexertion when pushing entities, falls, entrapment or being hit by a falling object (GANGULI; MILLER; POTHINA, 2021). After preprocessing the text data similarly to what was described below, the authors classified the events with an accuracy of up to 96%.

As for work-related diseases, it is possible to find some more regionalized studies, and in smaller numbers when compared with studies on occupational accidents. Noia et al. (2020), for example, evaluated the risk of work-related diseases in Italy for six types of pathologies: noise-induced hearing loss, spinal diseases, musculoskeletal diseases, tumors of the pleura and peritoneum, carpal tunnel syndrome, and skin diseases. Employing the SVM and KNN algorithms, in addition to using the unsupervised k-means algorithm as a "blind classifier" as a baseline, the authors divided the dataset into the mentioned classes and calculated the probability of occupational risk in each one of them (NOIA et al., 2020). Lu et al. (2019) used regression algorithms and time series analysis techniques to predict the number of occupational accidents in China, applying a combination of gray models with KNN, SVM, random forest, GBM, and ANN models. In India, Sau and Bhaktab evaluated the probability of workers in maritime activities

acquiring anxiety and depression (SAU; BHAKTA, 2019). They classified maritime workers into four classes (healthy, with anxiety and without depression, without anxiety but with depression, or with anxiety and depression), obtaining an 82.6% accuracy and a precision of 84.1% through the use of the Catboost algorithm.

In Table 2, we summarize the work developed using predictive models to solve problems involving occupational accidents and work-related diseases. We list the type of problem studies, the data used, the algorithms trained, and the results.

Reference	Type of ML problem	Data used	Algorithms
(CHARAPAQUI-MIRANDA et al., 2019)	Classification	Fit for work occupational health assessments	DT, RF, LR, SVM
(SARKAR et al., 2019)	Classification	Accidents occurred in India from 2010 to 2013	SVM and ANN; GA and PSO
(RECAL; DEMIREL, 2021)	Classification	Construction work-related accidents in Turkey	LR, SVM, ANN and SGB
(KHAIRUDDIN et al., 2022)	Classification	OSHA public accident data	SVM, KNN, NB, Decision Tree and RF
(KOC; EKMEKCI-OĞLU; GURGUN, 2021)	Classification	Construction work-related accidents in Turkey	RF, XGBoost, AdaBoost and Extra Trees
(SCOTT et al., 2021)	Classification	Pre-hospital care reports in USA	NB
(KOKLONIS et al., 2021)	Classification	Post-accident data collected at a hospital in Greece	NB, Bayesian Network, KNN, MLP
(GHOLAMIZADEH et al., 2023)	Regression	Occupational accidents in Iran	RF, SVM, MARS and M5
(TOLEDO; TIMOTEO; BARBOSA, 2020)	Classification	Work-related accidents in Brazil	LR, KNN, RF, LightGBM
(CHENG; KUSOEMO; GOSNO, 2020)	NLP + Classification	OSHA data on construction work-related accidents	KNN, SVM, ANN
(GANGULI; MILLER; POTHINA, 2021)	NLP + Classification	MSHA data containing descriptions of accidents in mining	RF
(NOIA et al., 2020)	Clusterization and Classification	Data on occupational illnesses in regions of Italy	K-means, SVM, KNN
(LU et al., 2019)	Regression	Work-related accidents in China	KNN, SVM, RF, GBM and ANN
(SAU; BHAKTA, 2019)	Classification	Data on maritime workers in India	Catboost, RL, NB, RF, and SVM

Table 2 – Summary of related works.

3 BRSTATS: OBTAINING A SOCIOECONOMIC STATISTICS DATASET OF THE BRAZILIAN CITIES

As a first step in this work, we proposed the construction of a unified dataset containing statistical data for all cities in Brazil. We use data acquired by some public organs: the Brazilian Institute of Geography and Statistics (IBGE, 2023e), the Institute of Applied Economic Research (IPEA)(IPEA, 2023), and the Brazilian Health Ministry(SAúDE, 2023). The data were extracted from public APIs (application programming interfaces) or through CSV (comma-separated values) files downloaded from the institutes' web pages, and then the tables were joined to obtain the final dataset.

It is noteworthy that several works in literature use socioeconomic indicators proposed in this work to explain and describe various phenomena, for example, the ones related to public health, agroindustrial production, and education in Brazil (FISCHER et al., 2007; SANTOS; BARBOSA, 2017; JAEN-VARAS et al., 2019) and in other countries (TANG et al., 2022; ZHANG et al., 2022; RODRÍGUEZ-RUEDA et al., 2021). Thus, the present work can also contribute to scientists and researchers in several fields, and the resulting dataset is publicly available in Zenodo, as informed in what follows.

It is also essential to observe that the features used in this work are periodically updated by the responsible institutes. It is worth calling attention to the fact that the sources used are not uniform, which has enlarged the effort in treating and integrating the data.

3.1 Sources and ETL process

In this section, we describe the sources used to build the BrStats dataset of Brazilian city statistics and the methods used to obtain it.

3.1.1 Data sources

The data sources used in this work are briefly described in the following section. It is worth calling attention to the fact that we only considered public sources available online, such that the development of this work finds no obstacle in Brazilian laws.

3.1.1.1 IBGE

IBGE is the main provider of statistical data in Brazil [1]. The main objective of the institute is to offer a complete view of the country through the production, analysis, and consolidation of statistical and geographic information.

[1] https://www.ibge.gov.br/acesso-informacao/institucional/o-ibge.html

The institute maintains an API called SIDRA (Sistema IBGE de Recuperação Automática) [2], from which one can obtain the data from its numerous surveys and research. To simplify the search and the use of SIDRA API, IBGE gives a corresponding code to identify each table and variable. Table 3 lists the table codes, variable names, and the corresponding variable code. We also inform the year of the last information on SIDRA.

Table code	**Table subject**	**Variable**	**Variable code**	**Last information**
6579	Resident population	Population	6579	2021
6449	Companies and other organizations	Working staff	707	2021
		Salaries	662	2021
		Companies	2585	2021
		Salaried staff	708	2021
1301	Surface area	Area	615	-
5938	Public finance	Gross Domestic Product	37	2020
5457	Agriculture	Cultivated area	8331	2021
		Harvested area	216	2021
		Agricultural production	215	2021
74	Livestock	Livestock production	215	2021

Table 3 – Description of the data extracted from IBGE.

While IBGE table n. 6579 brings the estimated population contingent for the Brazilian cities, table n. 6449 has information obtained in a continuous statistical pol named "Cadastro Central de Empresas", in which the institute obtains data about employment in the country (like the number of companies, total employed persons, salaried employed persons, and wages). Tables n. 1301 and 5938, respectively, bring information about the surface area of Brazilian cities and their Gross Domestic Product (GDP). Finally, tables n. 5457 and 74 deal with agriculture and livestock production in the country, bringing data about cultivated and harvested area and total production.

3.1.1.2 IPEA

IPEA is a Brazilian public institution that provides support to the federal government regarding fiscal, social, and economic public policies [3]. The institute publishes more than 250 studies annually, aiming to improve the efficiency of government decisions and, as a consequence, help the social, economic, and structural country's development.

The institute also provides a public API to simplify the data extraction[4], which was used in this work. We summarize, in Table 4, the data extracted from the API, listing the name of the table, the variable, and the year of the last information.

[2] https://apisidra.ibge.gov.br/

[3] https://www.ipea.gov.br/portal/categorias/110-conheca-o-ipea/13764-who-we-are

[4] http://www.ipeadata.gov.br/api/

Table name	Variable	Last information
EXPORTACAO	Exports	2021
IMPORTACAO	Imports	2021
RECTOTCH	Revenue	2021
RTRCORTOM	Current transfers	2021
RTRKTOM	Capital transfers	2021

Table 4 – Description of the data extracted from IBGE.

The tables named EXPORTACAO and IMPORTACAO bring information about the foreign trade of Brazilian cities, which is correlated with their economic production and welfare. Tables RECTOTCH, RTRCORTOM, and RTRKTOM are related to municipal finance, informing, respectively, the cities' total revenue, current and capital transfers. The last two variables are part of the balance of payments of public accounting and, while the capital transfers are related to the changes in changes in cities' ownership of assets, the current transfer quantifies the net incomes in the public sector (BANDY, 2018).

3.1.1.3 DATASUS

The Brazilian Health Ministry maintains a web application called DATASUS, which allows us to extract public data[5]. The data was downloaded directly from the web page in a CSV file.

In this work, we considered the information about born children and child death, such that we can estimate child mortality in Brazilian cities. The correlation between child mortality and various socioeconomic variables has been studied in the literature (FISCHER et al., 2007). So, this quantity can be related to the health condition of a city and, as a consequence, it helps to measure the human development of a locality (FISCHER et al., 2007). Therefore, it is an important feature to compose the dataset obtained in this work.

3.1.2 Extract, transform and load (ETL)

In Fig 1, we depict the extraction and transformation process of the data used in this work, which was developed using the Python programming language (ROSSUM; JR, 1995).

In the left part of the figure, we illustrate the extraction of data from multiple sources. As discussed previously, while the IBGE data is extracted from the Sidra web API[6] and the IPEA data is obtained through Ieadata API[7], the data acquired from DATASUS is downloaded in CSV files.

[5](https://datasus.saude.gov.br/informacoes-de-saude-tabnet/)
[6]https://apisidra.ibge.gov.br/
[7]http://www.ipeadata.gov.br/api/

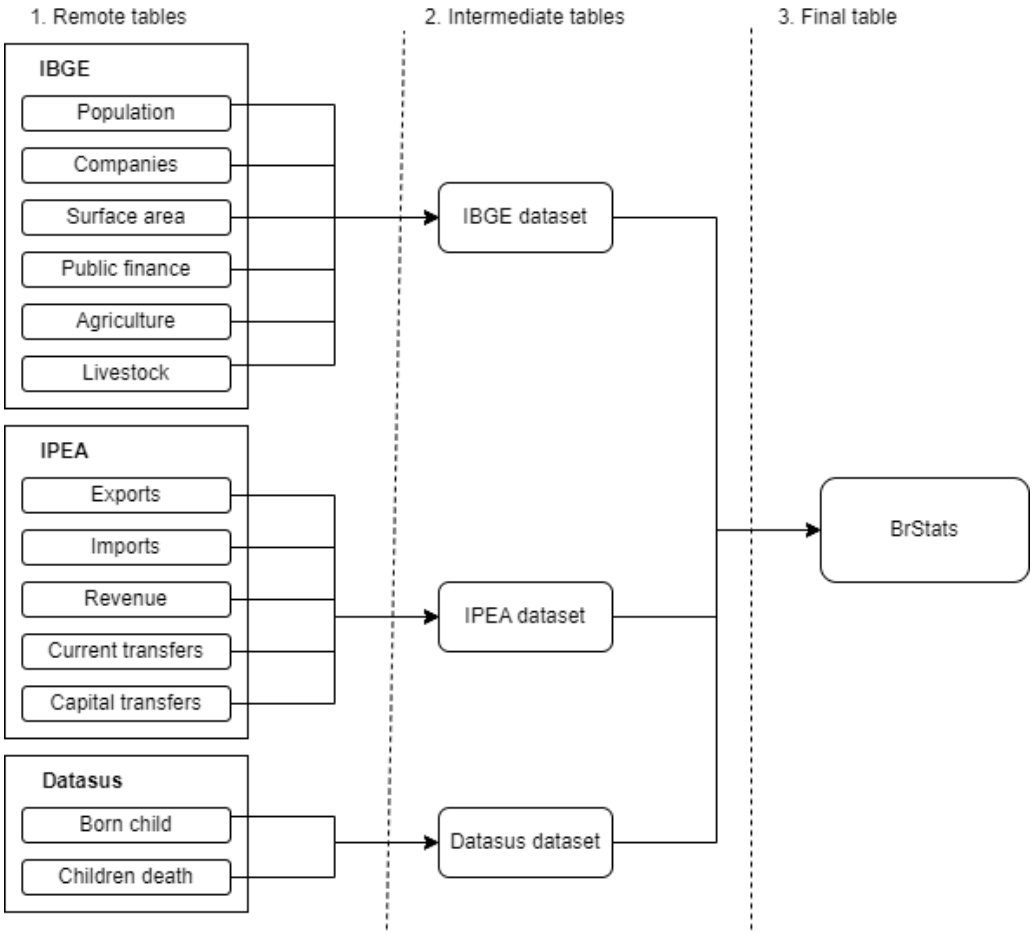


Figure 1 – Process of extraction and integration to obtain the resulting dataset.

After the extraction process, the tables have been aggregated by their origins: the first group is the data coming from IBGE, the second group is the data from IPEA, and the third one represents the public health data.

Finally, the three tables were integrated to obtain the final dataset. In this stage, some treatment must be executed due to the lack of standardization of data from different sources. It is necessary to mention that the IBGE adopts a numeric code for each city in Brazil, which contains seven digits, the last one being a check digit. The data coming from the DATASUS only contains six digits, so we need to unify the data using a unique code for each city.

3.2 The resulting dataset

After the ETL process described in Sec. 3.1, we obtain a dataset aggregating statistics for all Brazilian cities from 2012 to 2021. The obtained dataset has 36,315 rows and 21 columns related to socioeconomic variables. The BrStats dataset is publicly available [8].

[8] through the addresses <https://zenodo.org/records/15807785?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6ImJjMTRjYWViLWRiNGYtNDIwMS04MmNhLTRkYWNmZGM0MzEyYyIsImRhdGEiOnt9LCJyYW5kb20iOiI4ZGeU2J0ac34fLOFOJdDCTAWKFSicNgX6JiYQtcrX4WZXGbr-xrmsLPb3_lRMgTdrmJrH5P9LhPxTnAJmtBU3WAuQ> and <https://drive.google.com/file/d/1HBI054CnCCAX7kzAmEUUH2olhix2eiqe/view?usp=sharing>

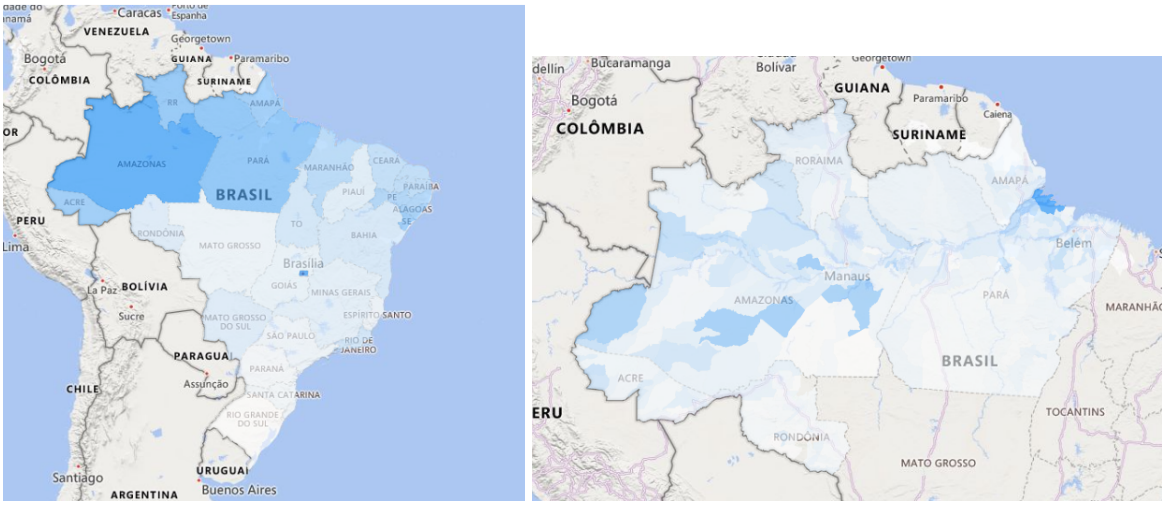


Figure 2 – Filled map of the salaried workers per company in Brazil states and in the cities of the Brazilian Northeast region.

3.2.1 Data dictionary

The data dictionary of the obtained dataset is represented in Table 5, in which we list the final table columns, the corresponding data type, the unit of measurement, the maximum and minimum values of the variables, and, finally, a synthetic description. All the variables are important indicators of the socioeconomic and demographic aspects of Brazilian cities.

It is important to notice that the IPEA data (exports, imports, revenue, current transfers, and capital transfers) are not available for all cities in the considered period, and thus, the BrStats user needs to observe the possibility of using or not using these data.

3.3 Examples and possible uses

The BrStats dataset can be used in a large number of proposes. Since the variables extracted may reflect the local culture and economy in Brazilian cities, they can be used to perform data analysis or as features in machine learning models.

Initially, we can observe that the dataset itself has a large number of variables that describe the socioeconomic profile of country regions. These variables can be used to generate dashboards and statistics of the municipalities or can be used to obtain an even larger number of consequent variables. In Fig. 2, for example, we calculated the number of salaried workers per company for all the cities and depicted this information in a filled map of Brazil. This information is an important signal of companies' automation and the use of personal labor, and these data can be used, for instance, in the distribution of labor inspections in the country.

Also, based on the BrStats content, we can analyze the profile of agricultural activity in Brazil, which is an important component of the country's economy. In Fig 3, we represent time series charts, respectively, for agricultural and livestock production for all Brazilian regions. We

Column name	Type	Unit	Min value	Max value	Null count	Description
Ano	int	-	2016	2021	0	year
CDMunicipio	str	-	-	-	0	IBGE code of the city
Populacao	int	-	771	1.23×10^7	30	City population
PessoalOcupado	int	-	20	7.32×10^6	30	Working staff
PessoalAssalariado	int	-	3	6.55×10^6	30	Salaried staff
VrSalarios	int	10^3 R$	35	3.21×10^8	30	Total salaries sum
PIB	int	10^3 R$	11679	7.63×10^8	5595	Gross Domestic Product
QtEmpresas	int	-	3	638246	30	Number of companies
AreaPlantada_h	int	$10^4 m^2$	0	9483	72	Cultivated Area
AreaColhida_h	int	$10^4 m^2$	0	9483	72	Harvested area
VlProducaoAgricola	int	10^3	0	9975	72	Agricultural production
VlProducaoPecuaria	int	10^3	0	8379	36	Livestock production
Area	float	km^2	3.6	159533.4	30	Surface area
Povoamento	float	person/km^2	0.03	14656.55	30	Nr. of people by km^2
Importacoes_US$	float	US$	1	1.52×10^{10}	20375	Total values of imports
Exportacoes_US$	float	US$	3	1.31×10^{10}	21392	Total values of exports
Receitas_R$	float	US$	9.36×10^6	6.48×10^{10}	11173	Total revenue
Transferencias_correntes_R$	float	R$	0	2.29×10^{10}	243	Current transfers
Transferencias_capital_R$	float	R$	0	8.37×10^8	243	Capital transfers
NrNascimentos	int	-	0	169299	0	Nr. of children born
NrObitosInfantis	int	-	0	1894	0	Nr. of children deceased

Table 5 – Data dictionary.

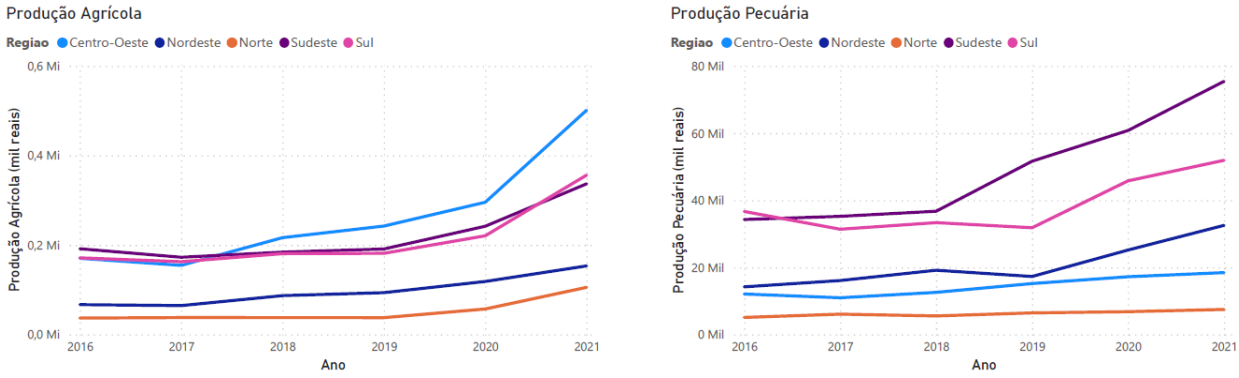


Figure 3 – Line graph for agricultural activity in Brazil.

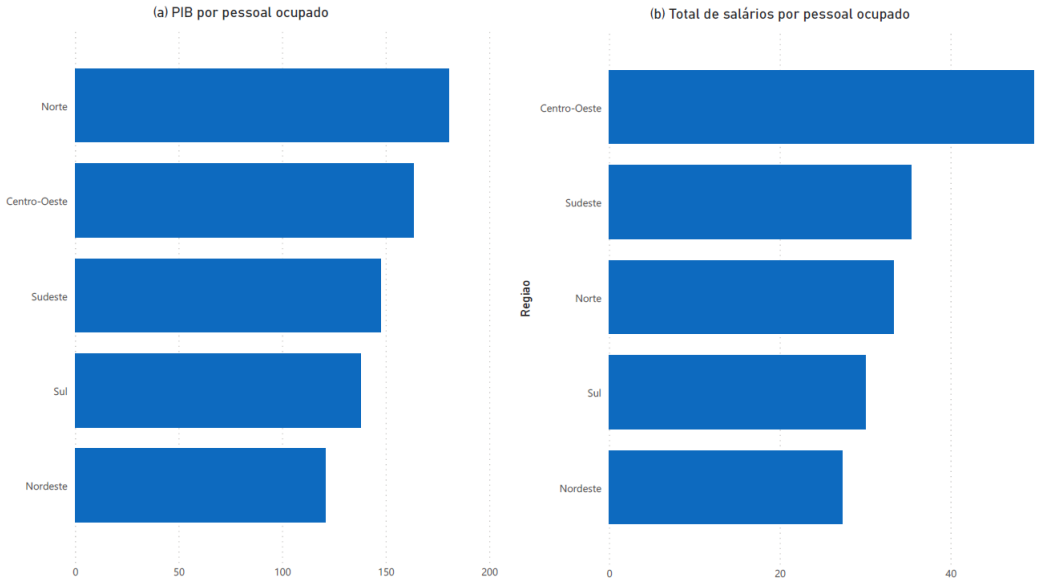


Figure 4 – Bar graphs for GDP per working person and salaries per working person in 2021.

can observe, for example, that the monetary earnings from agriculture in the Midwest region have grown in past years and become the highest in Brazil. On the other hand, the Southeast and South regions are the largest livestock producers in the country.

In Figure 4, we represent bar graphs for GDP per working person and mean salary per working person in Brazilian regions in 2021. We can observe that while the mean wage per laborer is higher in the Midwest and Southeast regions, the GDP per working person is higher in the North and Midwest regions. We can state that wages are lower in the North region, but it comprises a large part of the region's GPD.

In addition to the examples presented, it is possible to carry out a large number of analyses from the BrStats, which are beyond the scope of this work.

As discussed, the variables obtained in BrStats can be used as features in machine learning models. Aiming to help improve the effectiveness of the public policies in Brazil, in future contributions, we intend to use BrStats variables, together with other features, to predict

labor diseases and accidents in the country.

Given all the above, we can state that the dataset obtained in this work can be used in a variety of projects and studies in multiple areas of knowledge that need systematic information on Brazilian cities.

3.4 Challenges, limitations, and perspectives

The necessity of data extraction from multiple non-standardized sources brings some challenges to the methodology described above, which, however, can be the subject of future works.

Since we proposed making a public dataset of up-to-date features of Brazilian cities available and using data from public sources, we require the data sources to be updated by the respective institutes. This fact can be impacted in certain situations, such as the COVID-19 pandemic, during which some data updates were delayed.

Another challenge comes from the extraction process. As we used data from multiple sources and more than one method (API and CSV files), the source layout or URL change can impact the initial process in the ETL. Thus, in future contributions, we will need to observe alterations introduced by the Brazilian statistics institutes used as sources and update the extraction of the variables.

As discussed, Brazil is a big country with many cities. This fact makes obtaining data in some small regions very challenging. Consequently, statistical organs do not include some cities in the research. This fact brings some null variables to the shared dataset, which users must deal with.

It is worth mentioning that this work does not exhaust the possibility of aggregating variables for Brazilian municipalities. On the other hand, a continued effort is to search for more public variables and to obtain derived features from the ones already obtained.

4 PREDICTING OCCUPATIONAL ACCIDENTS IN BRAZILIAN STATES

In this chapter, we analyze the use of machine learning algorithms to predict the number of occupational accidents in each economic activity and Brazilian state

4.1 Exploratory data analysis

As an initial step in building predictive models, it is necessary to understand the data used as features and target variables. To this end, an exploratory analysis of the occupational accident data in Brazil is performed in this section.

In Brazil, all companies in which occupational accidents and work-related diseases occur are required to communicate these facts through a digital document named Occupational Accident Communication (CAT - Comunicação de Acidente de Trabalho). This communication has an important set of information about the employee (professional activity, age, and gender), the accident/disease (type of accident/disease, causative factor, etc.) the employer (such as location and economic activity). If the employer does not report the occupational accidents, the employees, the labor unions, the health system, or the government can report the CAT data. The Brazilian government receives and stores this data, which is used to create public workplace safety policies. This occupational accident dataset is also used in this work.

It is important to mention that we do not consider work-related diseases and maintain only occupational accidents in Brazil. From 2016 to 2022, a total of 2.387.938 occupational accidents were reported in the country, which will be analyzed in what follows.

We represent the line plot of occupational accident numbers in Brazil in Fig. 5 (blue line) and the deaths resulting from these accidents (red line) for the considered period. Due to the COVID-19 pandemic outbreak, the number of accidents and deaths decreased in 2020, maintaining levels of 450,000 (four hundred and fifty thousand) annual accidents and more than 2 thousand deaths.

In Fig. 6, we show the distribution of occupational accidents in Brazil by sex and age of workers. In Fig. 6a, the pie chart of the occupational accidents by sex is depicted, while we show in Fig. 6b the age pyramid of these accidents in Brazil. In the country, young men aged between 21 and 25 years are the most affected group by occupational accidents. We can also observe that the number of occupational accidents is higher among men (69.4% of occupational accidents occur among men). This demographic distinction may be explained by the professional activity type carried out by male workers in Brazil and the inexperience of young people at work.

We depict the distribution of occupational accidents by the type of injury in Fig. 7. We

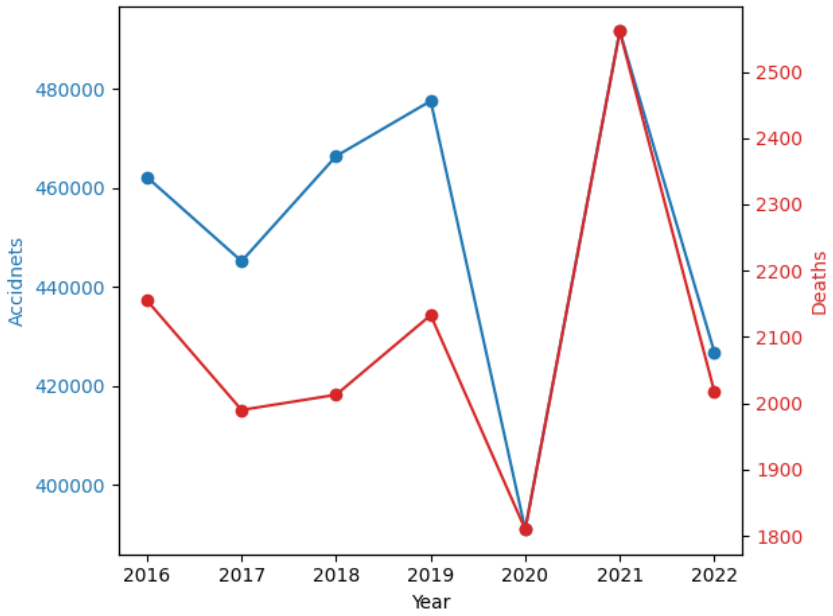
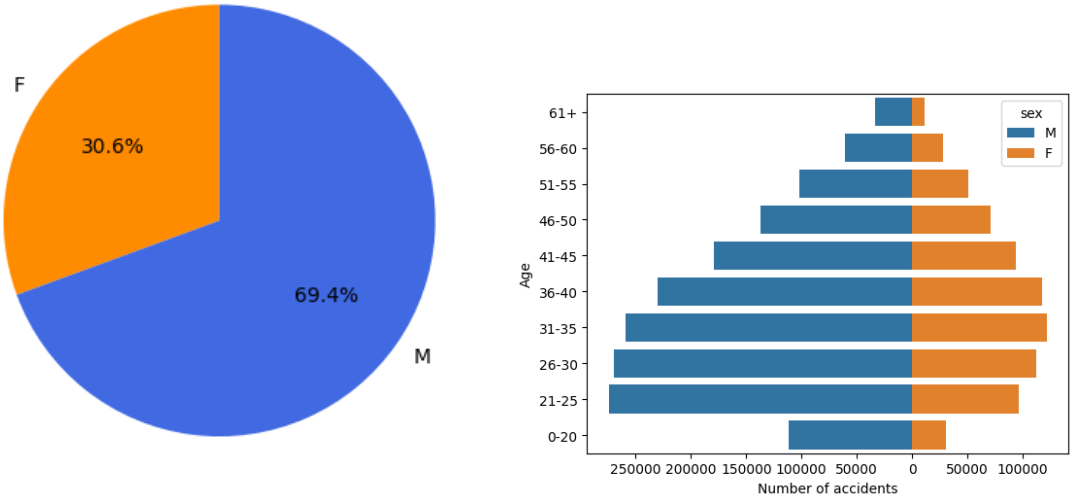


Figure 5 – Line plots of the number of occupational accidents and work-related deaths. Reprinted from (TOLEDO; MOURA, 2024).



(a) Pie chart: Work-related diseases by sex. (b) Age pyramid of work-related diseases in Brazil.

Figure 6 – Work-related diseases by sex and age. Reprinted from (TOLEDO; MOURA, 2024).

use the International Statistical Classification of Diseases (ICD) in this graph and show only the ten most frequent types of injuries. The most frequent injuries in the country are those related to musculoskeletal factors (such as hand and wrist injuries and fractures and foot and ankle injuries). Communicable diseases are also on the list, mostly related to health assistant professionals and influenced by the COVID-19 pandemic outbreak.

The professional activity executed by the employee can also be a determining factor in the occurrence of occupational accidents. In Brazil, we use a classification called the Brazilian Classification of Occupations (CBO) to categorize professional activities, to reflect the reality of

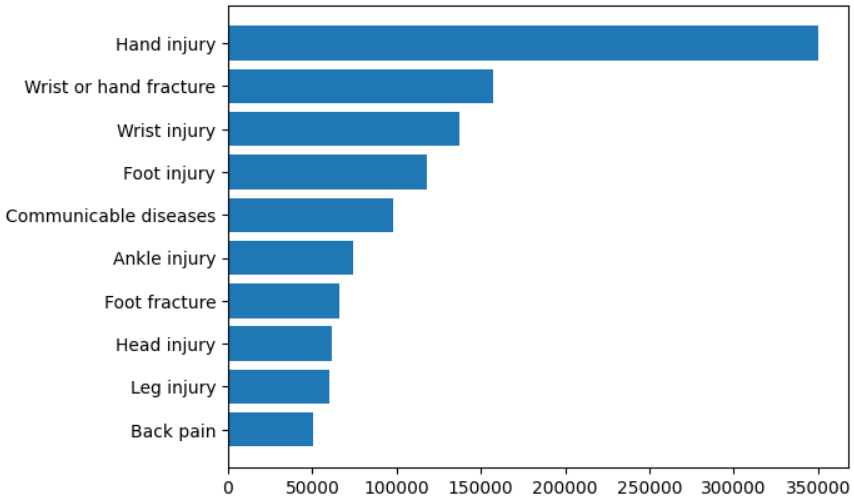


Figure 7 – Bar diagram of the distribution of occupational accidents in Brazil by type of injury for the ten most frequent types. Reprinted from (TOLEDO; MOURA, 2024).

professions in the Brazilian job market, and to monitor the dynamism of occupations. In Fig 8, we represent the number of occupational accidents per CBO family for the ten most frequent classes.

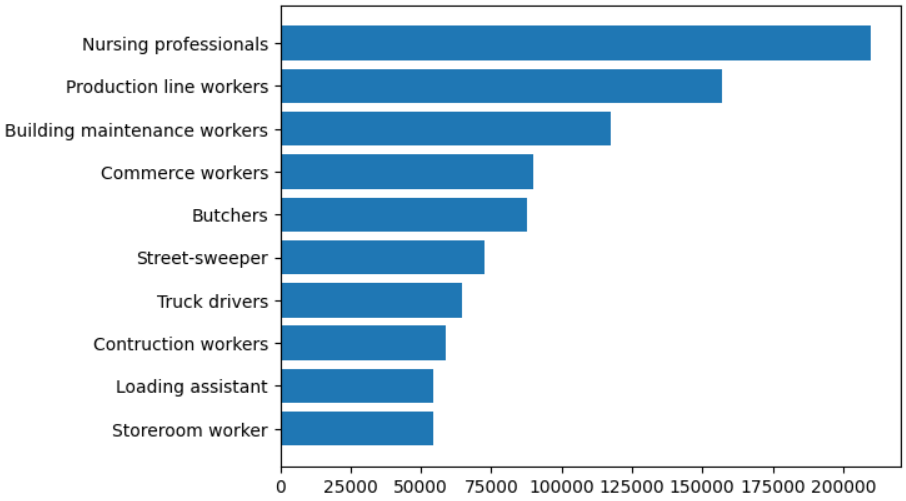


Figure 8 – Bar diagram of the distribution of the work-related diseases in Brazil by type of worker for the ten most frequent classes.

Nursing professionals appear at the top of the list, possibly because of the greater reporting of occupational accidents by hospitals and other health employers. Activities that stand out in the list as workers in the industry workers, building maintenance workers, and butchers work in activities with various associated occupational risks, which favor the occurrence of accidents.

Since we intend to predict the number of occupational accidents in Brazilian states, the target variable is obtained from the CAT dataset. The exploratory analysis performed in this section is also important to select the independent variables, as discussed in Sec. 4.2.

4.2 PROPOSED APPROACH

We analyze the methodology used in this work, describing in detail the extract, transform, and load (ETL) process to obtain the dataset used in the subsequent ML model training.

4.2.1 The methodological path

In Fig. 9, we summarize the methodology adopted in the present study. We can divide this study into three main steps: data preparation, data pre-processing, and ML model training and evaluation. In the first step (data preparation), we execute an ETL process to integrate data from multiple sources and obtain a single dataset with the features and the target variable, as described in this section. Then, we execute a preprocessing step and split the dataset into a training and a test one, as described in what follows. Finally, we train the ML models using the training dataset and evaluate the results in the test data, as shown in Fig. 9. The data preprocessing and ML model training end evaluation are described in Sec. 4.3.

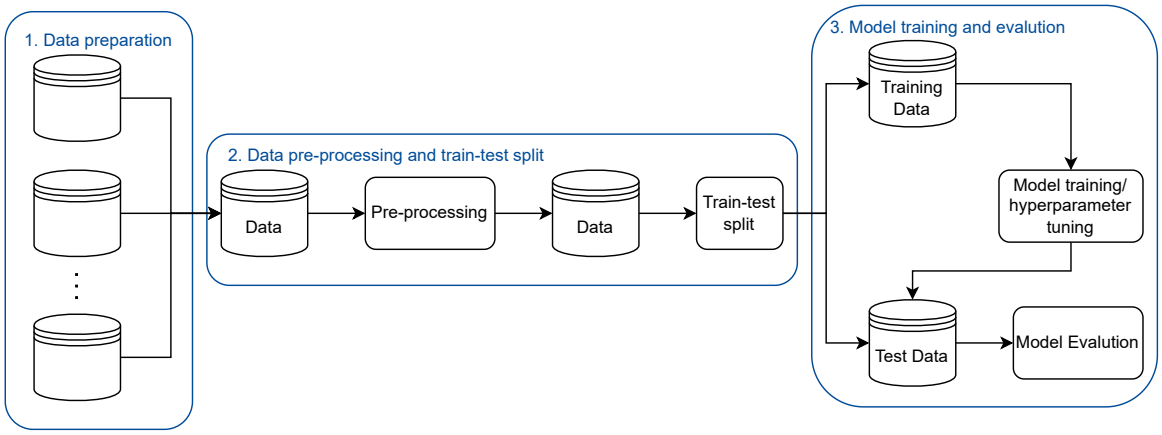


Figure 9 – The methodological path used in this work. Reprinted from (TOLEDO; MOURA, 2024).

It is important to highlight that we used the Python programming language (ROSSUM et al., 2007) in all the steps performed in this work and described in Fig. 9.

4.2.2 Data preparation

In this section, we aim to predict the number of occupational accidents in Brazilian states in each economic activity and by year. Thus, we use the mentioned CAT communication data to obtain the number of occupational accidents in the proposed granularity. It is important to note that we preserved data corresponding to the period between 2016 and 2021 in the current study since we have all the features available for the period.

An important contribution of this work is the ETL process shown in Fig. 9, from which multiple datasets are integrated to obtain a single dataset containing all variables described below.

The source datasets are both public data maintained by Brazilian statistics agencies and the Brazilian Labor Ministry databases.

Using the public sociodemographic Brazilian datasets, Toledo and Moura have obtained a single dataset containing data for all the country cities (TOLEDO; MOURA; TIMOTEO, 2023). This dataset includes data related to population, economy, employment, education, and health (TOLEDO; MOURA; TIMOTEO, 2023). In this work, we employ some of these variables to compose the ML training dataset.

We include in the features set two population statistics variables: the city population and working staff. We can expect that the bigger the population and working staff, the greater the number of occupational accidents in the city.

We also may expect that the occurrence of occupational accidents is related to the level of economic activity in a region and to its human development. In this view, we include as features the Gross Domestic Product (GDP), which calculates the market value of all the finished goods and services produced in a region, and the Human Development Index (HDI), which measures the indicators related to health, education, and work conditions, key dimensions of human development.

Variables related to employment in the country were obtained from the Brazilian Labor Ministry databases and included in this study. From this source, we acquire the number of employers and employees in each Brazilian state. We also include as a categorical variable the Brazilian National Classification of Economic Activities (CNAE), which describes the activity developed by an enterprise in the country. We calculate and include as features the proportion of female workers, the mean age of the employees, and the average time they work with a given employer since these were important statistical factors in the occurrence of occupational accidents, as shown in Sec. 4.1.

Finally, we include features obtained from the Brazilian Labor Inspection. The WHO/ILO joint estimates of the work-related burden of disease and injury point out exposure to long work hours as the major cause of deaths related to work and informal jobs being correlated to the occurrence of accidents (ORGANIZATION et al., 2021). Thus, we include as variables the number of irregularities related to informal workers and the number of irregularities related to working hours. Brazilian legislation requires Labor Inspectors to stop work activities if serious and imminent risks to workers' health are detected, in procedures called embargos or interdictions. The numbers of these procedures per economic activity in a given state were also included as features.

It is worth mentioning that the described datasets are joined using the Brazilian cities and the year as keys.

4.2.2.1 The resulting dataset

Integrating the datasets described below, we obtain a unified dataset containing the number of occupational accidents and all the other variables, for each Brazillian city. These cities are grouped into 27 states that are used as territory granularity in this work. Thus, we aggregate the described variables, summing the numerical variables or calculating the mean value for the ones that are average numbers. At this stage, we calculate the population density (ratio between population and surface area) and employers' density (number of employers divided by the surface area). In Table 6, we show the features used in this work, describing, for each variable, its type, unity, maximum and minimum values.

Variable	Description	Type	Unit	Min value	Max value
UF	Brazilian state	string	-	-	-
Cnae	Brazilian economic activity classification	string	-	-	-
Population	Population	int	-	1.85×10^3	3.08×10^7
WorkingStaff	Working staff	int	-	0	2.33×10^6
PopulationDensity	Number of people by km^2	float	-	0.6	5363.08
HDI	Human Development Index (HDI)	float	-	0.469	0.847
GPD	Gross Domestic Product (GDP)	float	10^3 R$	2.36	2.18×10^7
NrEmployers	Number of employers	int	-	0	633,656
EmployersDensity	Number of enterprises by km^2	float	-	0	5.36
NrEmploees	Number of employees	int	-	0	1.07×10^6
PropFemale	Proportion of female workers	float	-	0	1
AvgAge	Average age of employees	float	years	0	56.70
AvgTime	Average time working for the employer	float	years	0	25.55
NrIrregularities	Nr. of irregularities related to informal workers	int	-	0	662
NrIrregHours	Nr. of irregularities related to working hours	int	-	0	270
NrEmbargoes	Nr. of embargoes/closures	int	-	0	1547

Table 6 – Data dictionary. Reprinted from (TOLEDO; MOURA, 2024).

We represent the correlation heatmap between numeric features in Fig. 10. It is essential to state that we take into account the correlation between variables when choosing the features for model training, and if a feature pair has a correlation near one, we removed one of them, maintaining only the ones listed in Table 6 and in Fig. 10. For example, initially, we intended to use the total number of female workers and the total salaried workers as features. But as the number of female workers and the number of employees have a Pearson correlation near 1, only the second variable is maintained. Similarly, the total salaried population and the working staff have a correlation coefficient near one and, thus, the first feature was removed.

4.3 EXPERIMENTAL PROTOCOL

We discuss the final steps presented in Fig. 9: the data prepossessing and the machine learning models' training.

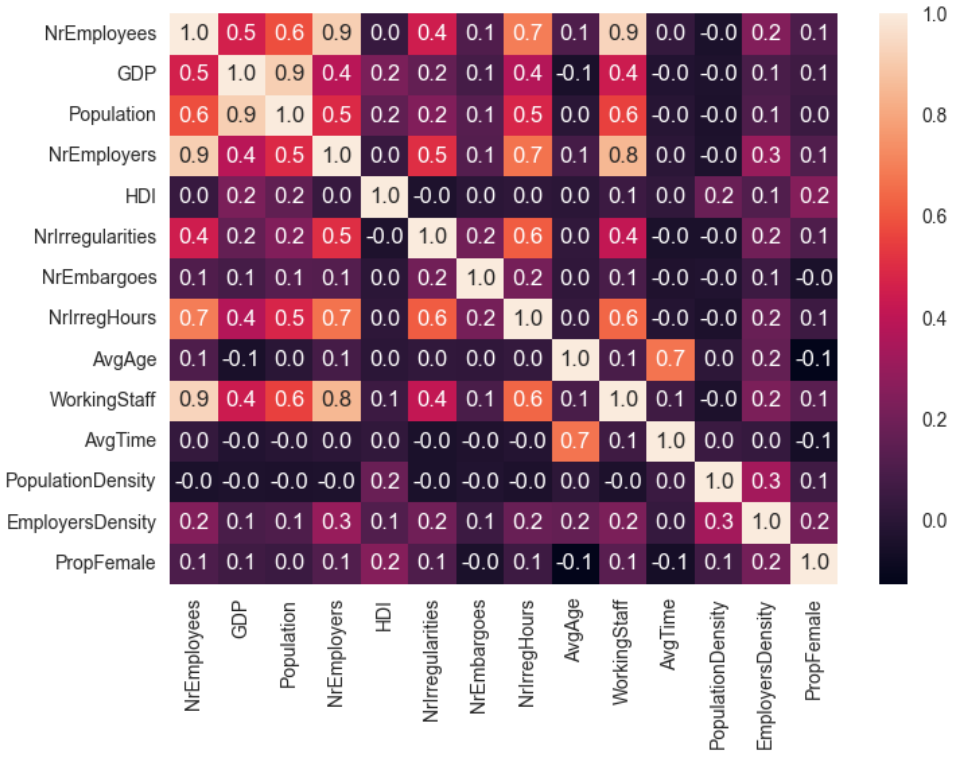


Figure 10 – Features correlation heatmap.

4.3.1 Data preprocessing

Before we train ML models, there is still an important task to be accomplished, which is to preprocess the data. In this view, firstly, all null numerical data are replaced by zero.

Since we use linear regression as a baseline model, the categorical variables must be transformed into numerical ones. In this work, we deal with a large number of variable categories and then adopt the target encoding strategy (MICCI-BARRECA, 2001), proven to be effective in similar problems (PARGENT et al., 2022). In this strategy, each category is encoded based on a shrunk estimate of the average target values for observations belonging to the category.

Finally, the numerical variables were standardized by the common standard scalar strategy: the variable values were subtracted from their mean and divided by their standard deviation.

After the ETL process and the preprocessing step, the resulting dataset has 11,255 (eleven thousand, two hundred and fifty-five) rows.

4.3.1.1 Train-test split

Dividing the dataset into train and test one is a common step used in ML. The first one is used to train the algorithm and adjust its parameters. The predictions are evaluated in the test dataset. In the current work, the dataset resulting from the preprocessing step is randomly divided into a training dataset (with 80% of the data instances, or 9,004 -nine thousand and four)

Model	Hyperparameter search space
SVR	C:[0.1,1,10,100] gamma:['scale', 'auto'] kernel:['linear', 'poly', 'rbf']
XGBoost	max_depth: [5, 9] max_leaves: [6, 17] booster: [gbtree, dart] subsample: [0.7, 0.8, 0.9, 1.0] colsample_bytree: [0.8, 0.9, 1.0] learning_rate: [0.05, 0.5]
LightGBM	max_depth: [5, 9] num_leaves: [6, 17] boosting_type: [gbdt, dart] subsample: [0.7, 0.8, 0.9, 1.0] colsample_bytree: [0.8, 0.9, 1.0] learning_rate: [0.05, 0.5]

Table 7 – Hyperparameter search spaces. Adapted from (TOLEDO; MOURA, 2024).

and a test dataset (with 20% of the data instances, or 2,251 - two thousand, two hundred and fifty one).

4.3.2 Moldel training

In this work, we evaluate the use of ML algorithms to predict the number of occupational accidents in Brazilian states, for each economic activity developed in the region. Since we predict a continuous real number, we need to make use of regression models.

The simplest regression algorithm, which has been known in the literature for centuries, is linear regression (JAMES et al., 2013). We use linear regression in this study as a baseline model to compare the results obtained from the more modern algorithms.

In this work, we apply models SVM and gradient boosting (XGBoost and LightGBM), since they have been winning many ML competitions and presenting high prediction performance (BENTÉJAC; CSÖRGŐ; MARTÍNEZ-MUÑOZ, 2021; NIELSEN, 2016). These models have also been used in problems similar to the one proposed in this work (NOIA et al., 2020; TOLEDO; TIMOTEO; BARBOSA, 2020).

The ML models have a set of hyperparameters, which can be set to improve the models' performance. We display in Table 7 the hyperparameter search space used for SVM and the gradient boosting models used 7.

As a means of choosing the best hyperparameters for the models, we use cross-validation with four folds and Bayesian optimization. The training dataset is divided into four folds, with three of them being used for training the model, while the last one is used for evaluating the predictions. After the process of iterations, the best hyperparameters are chosen, and the whole

Model	R^2	MAPE	RMSE
Linear regression	0.492	21.27%	743.20
SVR	0.725	3.31%	546.54
XGBoost	0.884	1.82 %	401.80
LightGBM	0.908	1.86 %	316.60

Table 8 – Metrics for the implemented regression models. Adapted from (TOLEDO; MOURA, 2024).

dataset is used to train the algorithm, which is evaluated with the test dataset.

4.4 RESULTS AND DISCUSSION

In this section, we discuss the results obtained by the ML models trained as shown in Sec. 3.3.

We present the metrics obtained for the models in the test dataset in Table 8. The gradient boosting algorithms have higher R^2 and lower values of $MAPE$ and $RMSE$. We can state that the features chosen and the models trained explain the independent variable (the number of accidents) since the high values of R^2 indicate the distancing of the prediction from random guess. This metric reaches the values of 0.884 for XGBoost and 0.908 for LightGBM, values superior to the one obtained for linear regression, the baseline model used in this work.

Analyzing Table 8, we can also observe that there is only a 1.86% percentual difference between the actual and predicted values of occupational accidents when using the LightGBM model and 1.82% for XGBoost. On the other hand, the values of RMSE are 401.80 for XGBoost and 316.60 for LightGBM, below 1051.72, which is the standard deviation of the number of occupational accidents (target variable).

In Fig. 11, we represent the prediction error plots for the trained models in the test dataset. On the horizontal axis of the graphs, we depict the actual values of the target variable, while on the vertical axis, we observe the predicted values for the variable. In these plots, the identity line represents a scenario where the prediction exactly matches the model. We can also observe how much variance exists in the predictions. We can notice in Fig. 11 that the Linear regression model has the higher variance while the LightGBM presented the lower one.

Gradient boosting algorithms calculate a score for each feature, representing the feature's importance, with a higher score representing a larger effect on the prediction. In Fig. 12, we depict the relative feature importance for the LightGBM model. The Brazilian state (UF), the total working staff, and the worker average time in the enterprise have a higher influence on the predictions. We can state that socioeconomic statistics variables (such as working staff, population density and HDI) have high importance, as depicted in Fig. 12.

We can observe that the Brazilian state has the highest importance. The territory is related to the economic activities developed and to the population, which can explain the score.

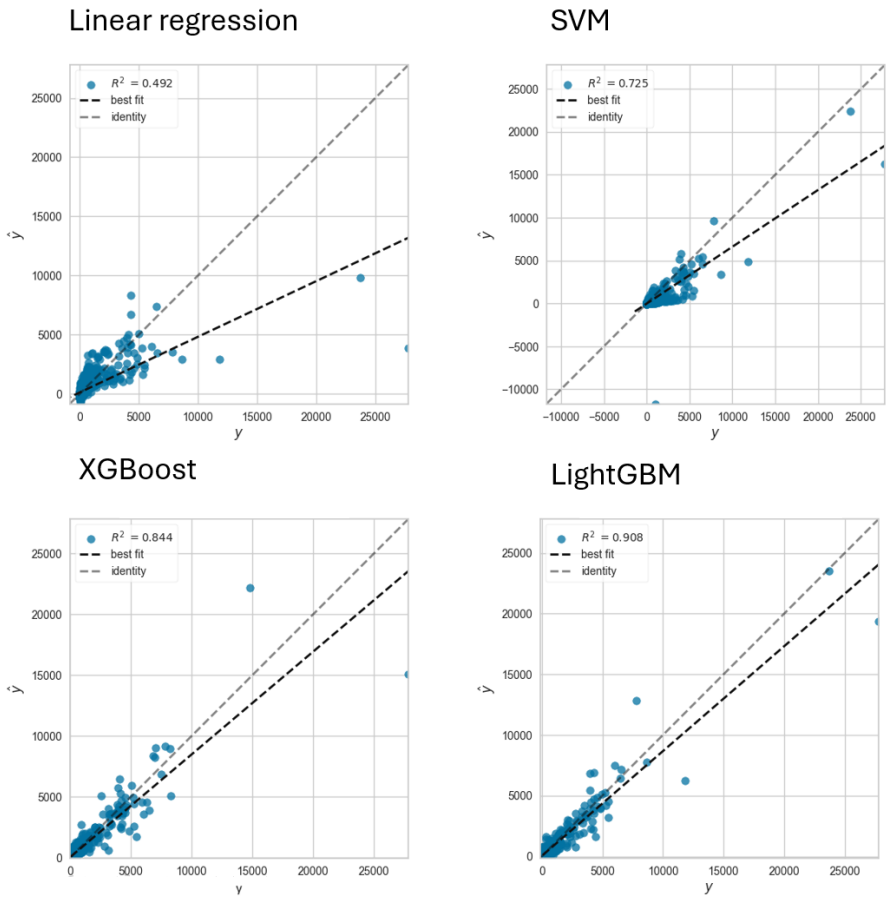


Figure 11 – Prediction error plots for the trained models in the test dataset.

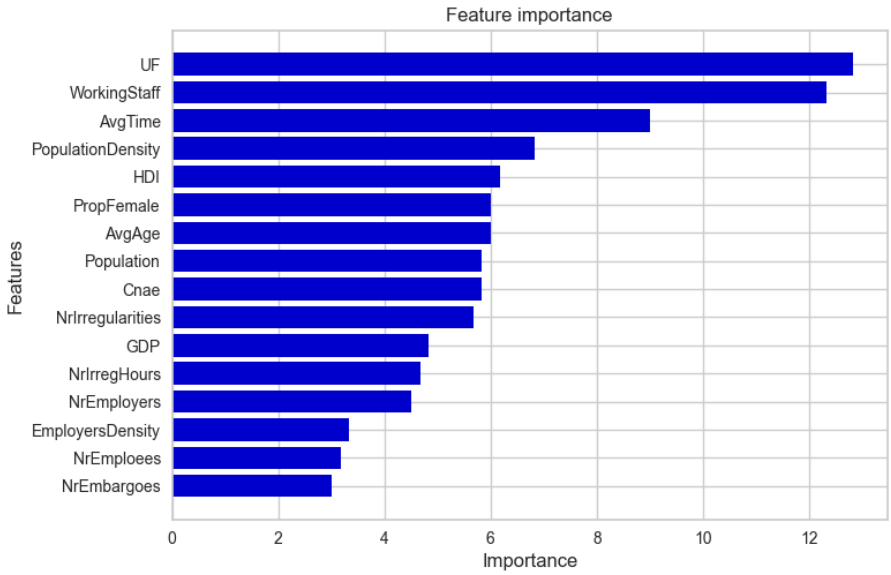


Figure 12 – Feature importance for LightGBM algorithm. Reprinted from (TOLEDO; MOURA, 2024).

Model	Best hyperparameters
SVR	C = 10 gamma='auto' kernel= 'poly'
XGBoost	max_depth = 5 max_leaves= 17 booster=dart subsample=0.7 colsample_bytree=0.8 learning_rate= 0.174
LightGBM	max_depth = 9 num_leaves= 7 boosting_type=gbdt subsample=0.7 colsample_bytree=1.0 learning_rate= 0.48

Table 9 – Hyperparameter search spaces. Adapted from (TOLEDO; MOURA, 2024).

The total work staff is the second most important feature since we can expect a growth in the number of accidents in territories with a higher number of workers. The average time that the employees work with the employers is also an important feature, indicating that the experience in the workplace reduces the probability of accidents.

Finally, for reproducibility reasons, we list in Table 9 the models' hyperparameters that gave the best metrics in the training step.

5 PREDICTING OCCUPATIONAL ACCIDENTS IN BRAZILIAN CITIES

In this chapter, we propose using machine learning (ML) algorithms to predict the number of occupational accidents in each economic activity in Brazilian cities. After obtaining all the variables used as features, we train ML models to produce consistent results and avoid overfitting, as described in what follows.

We use linear regression as a baseline model and train Decision Trees (DT) and gradient boosting algorithms (GradientBoosting, LightGBM, XGBoost, and CatBoost) to predict the number of occupational accidents in each economic activity and each municipality in Brazil.

We also aim to evaluate which variables are most important in building the models and how the choice of these variables influences the predictions made. To this end, we developed an automated feature selection strategy to list the variables by their importance and, iteratively, remove the less important features, and train the models, evaluating the results obtained. We also use hypothesis tests to check if there may be a real difference between the algorithms' predictions.

5.1 METHODOLOGICAL PATH

In Fig. 13, we represent the methodological path adopted in this work. Firstly, we extract data from multiple sources (data on occupational accidents in Brazil, Brazilian statistical dataset, data on employment, and labor inspection data), obtaining an integrated dataset used in the following steps. As discussed below, we perform a preprocessing step, and the resulting dataset is split into a training dataset and a test one. We use the training dataset to train the ML model, choose the models best hyperparameters, and iteratively select the most important features. Thus, the test dataset is used to evaluate the predictions.

The steps presented in Fig. 13 are analyzed throughout the text. In subsection 5.1.1, we discuss the data extraction and integration, while, the pre-processing step and ML models training are analyzed in Sec. 5.3.

5.1.1 Extract, transform and load - ETL

The first step shown in Fig 13 is the extraction of data from multiple sources, and integration to obtain a unified dataset used to train ML models. Let us detail this fundamental step in this work.

According to Brazilian law, all companies must inform the government of all occupational accidents and work-related diseases through a digital document called Occupational Accident

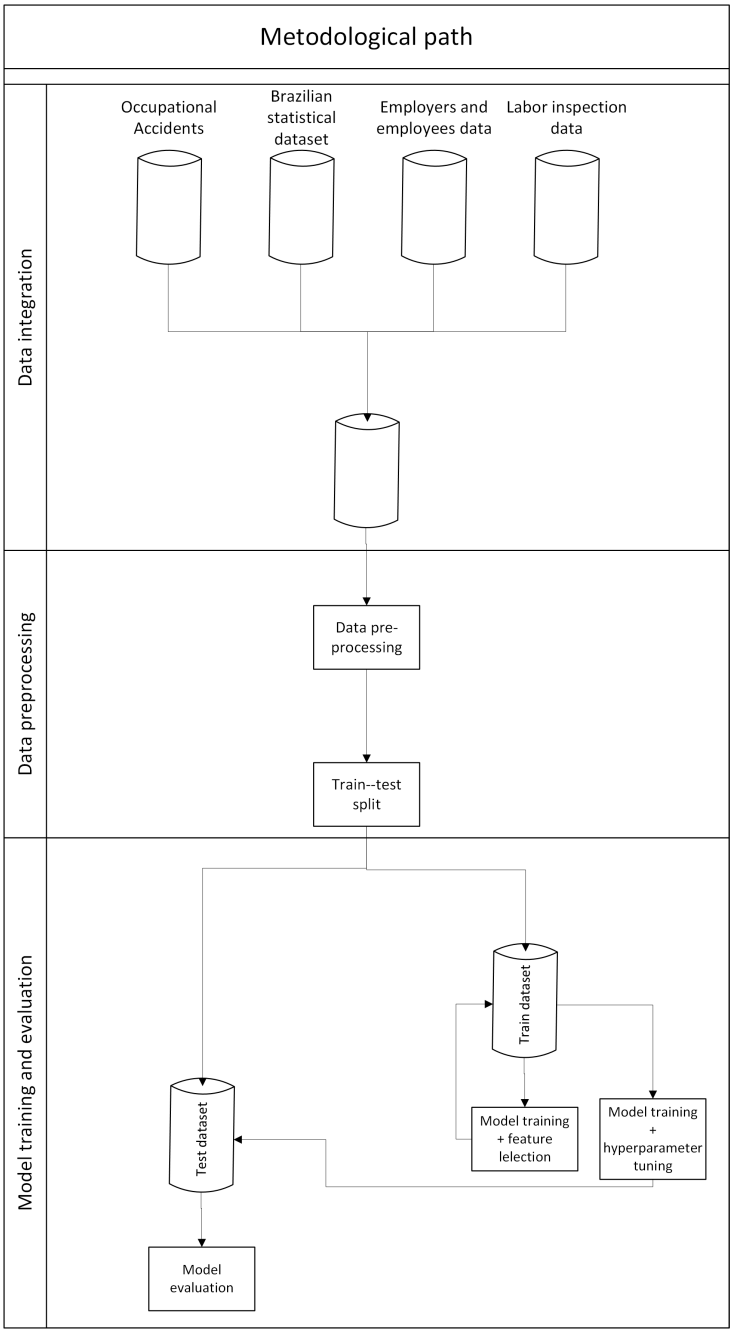


Figure 13 – The methodological scheme adopted in this work.

Communication (CAT - Comunicação de Acidente de Trabalho) and, from 2016 to 2022, more than 2.3 million occupational accidents were reported in Brazil. This document contains information about the employee (like age, gender, and professional activity), the accident/disease (type of accident/disease, causative factor, etc.), and the employer (such as its economic activity). The Brazilian government stores these data in a private dataset, which is used in this work. It is worth mentioning that the data is anonymized in such a way that no personal information (like name) is used.

Brazil has some statistical institutes that perform studies related to population, economy, employment, education, and health. Among them, we can mention the Brazilian Institute of

Geography and Statistics (Instituto Brasileiro de Geografia e Estatística - IBGE in Brazilian Portuguese)[1] and the Institute of Applied Economic Research (Instituto de Pesquisa econômica aplicada - IPEA)[2]. These institutes make the results of their research available through multiple public APIs (application programming interfaces)[3,4]. Although it is not simple to integrate information from these sources, a unified dataset with statistical data for all cities in Brazil was already created (TOLEDO; MOURA; TIMOTEO, 2023). In this work, we extend the work of Toledo et al. (TOLEDO; MOURA; TIMOTEO, 2023), incorporating variables into the dataset already obtained.

From the surveys conducted by IBGE, we obtained data related to cities' populations, working staff, and salaried staff (TOLEDO; MOURA; TIMOTEO, 2023). We can expect that the bigger the population and working and salaried staff, the greater the number of occupational accidents in the city. We also obtain the population density by the ratio between a city's population and its area.

The economic variables of the country's cities can influence the labor markets and, consequently, the occurrence of occupational accidents. Therefore, we add to this work some variables obtained from the surveys developed by IPEA. This public Brazilian institution provides fiscal, social, and economic data to improve the efficiency of government decisions. We include municipal revenues, total imports, and exports as features. We also include the values of cities' current income and capital income.

We may expect that the occurrence of occupational accidents is related to the level of economic activity in a region and to its human development. In this view, we include as features the Gross Domestic Product (GDP), which calculates the market value of all the finished goods and services produced in a region, and the Human Development Index (HDI), which measures the indicators related to health, education, and work conditions, key dimensions of human development.

Variables related to employment in the country were obtained from the Brazilian Labor Ministry databases and included in this study. From this source, we acquire the number of employers and employees in each Brazilian state. We also include the Brazilian National Classification of Economic Activities (Cadastro Nacional de Atividades Econômicas - CNAE) as a categorical variable, which describes the activity developed by an enterprise in the country. We calculate and include as features the proportion of female workers, the mean age of the employees, and the average time they work in a given employer, since these are important statistical variables in the occurrence of occupational accidents, as analyzed in Sec. 5.2 and already studied in literature (TOLEDO; MOURA, 2024). Once more, it is important to mention that all the personal data is anonymized in this work.

[1] https://www.ibge.gov.br/acesso-informacao/institucional/o-ibge.html
[2] https://www.ipea.gov.br/portal/categorias/110-conheca-o-ipea/13764-who-we-are
[3] https://apisidra.ibge.gov.br
[4] http://www.ipeadata.gov.br/api/

Finally, we add features obtained from the Brazilian Labor Inspection. We include the number of fines applied by Labor Inspectors and the number of irregularities related to informal workers. Since exposure to long work hours is one of the major causes of deaths related to work (ORGANIZATION et al., 2021), we include as a feature the number of irregularities related to working hours. Brazilian legislation requires Labor Inspectors to stop work activities if serious and imminent risks to workers' health are detected, in procedures called embargos or interdictions. The numbers of these procedures per economic activity in a given city were also included as features.

5.1.2 The dataset

The tables obtained from the occupational accidents dataset, Brazilian statistical dataset as described in (TOLEDO; MOURA; TIMOTEO, 2023), employment data, and labor inspection data have the Brazilian city and the year as common variables. These two columns are, then, used to join the tables.

After integrating the datasets, we obtain a unified dataset containing the number of occupational accidents and all the other variables for each Brazilian city. Tab. 10 describes the features present in the dataset, listing, for each variable, its type, unit, maximum and minimum values.

Some comments on the variables listed on Tab. 10 still need to be made. Firstly, the UF and the city code are categorical features related to the Brazilian state and the municipality. Concerning population data, while working staff corresponds to people who work in any activity, the salaried staff, a subset of working staff, is the sum of employees in a region. It is important to point out that, in this work, we analyze the influence of similarity between features like the ones mentioned in the models' predictions.

In Tab. 10, we can also observe that we maintain not only the HDI index but also its key dimensions: the HDI health (assessed by life expectancy at birth), HDI education (measured by mean of years of schooling for adults aged 25 years and more and expected years of schooling for children of school entering age), and HDI standard of living (measured by gross income per capita). In model training and evaluation steps, we analyze which features have to be maintained to increase the model's performance.

5.2 Exploratory Data Analysis

As an initial step in building predictive models, it is necessary to understand the data used as features and target variables. Thus, let us perform an exploratory analysis of the dataset obtained in the previous section.

Brazil has 5,570 (five thousand, five hundred and seventy) cities organized into 26 (twenty-six) states and one federal district. As already mentioned we intend to predict the number

Variable	Description	Type	Unit	Min value	Max value
City	Brazilian city	string	-	-	-
UF	Brazilian state	string	-	-	-
Cnae	Brazilian economic activity classification	string	-	-	-
Population	Population	int	-	812	1.1×10^7
WorkingStaff	Working staff	int	-	0	9.85×10^5
SalariedStaff	Salaried staff	int	-	0	8.56×10^5
PopulationDensity	Number of people by km^2	float	-	0.13	303.5
HDI	Human Development Index (HDI)	float	-	0	0.862
HDIh	HDI health	float		0	0.894
HDIe	HDI education	float	-	0	0.825
HDIs	HDI standard of living	float	-	0	0.891
GDP	Gross Domestic Product (GDP)	float	10^3 R$	2.36	7.14×10^8
NrEnterprises	Number of enterprises	int	-	0	303,772
NrEmployers	Number of employers	int	-	0	32,3006
EmployersDensity	Number of enterprises by km^2	float	-	0	0.82
NrEmploees	Number of employees	int	-	0	1.07×10^6
NrFemales	Number of female employees	int	-	0	281473
PropFemale	Proportion of female workers	float	-	0	1
AvgAge	Average age of employees	float	years	0	56.70
AvgTime	Average time working for the employer	float	years	0	25.55
NrIrregularities	Nr. of irregularities related to informal workers	int	-	0	211
NrIrregHours	Nr. of irregularities related to working hours	int	-	0	93
NrEmbargoes	Nr. of embargoes/closures	int	-	0	978
NrFines	Nr. of labor fines	int	-	0	2722
Revenues	Municipal revenues	float	US$	0	5.43×10^{10}
CurrentIncome	Current income	float	US$	0	$1{,}85 \times 10^{10}$
CapitalIncome	Capital income	float	US$	0	8.37×10^{10}
Exports	Total of exports	float	US$	0	1.31×10^{10}
Imports	Total of imports	float	US$	0	1.52×10^{10}

Table 10 – Data dictionary.

of occupational accidents for each economic activity of each Brazilian town. Thus, we show the ratio distribution between accidents and the working personnel in each Brazilian city in Fig. 14. This map shows that occupational accidents occur in all Brazilian cities, but are more concentrated in the largest municipalities like the states' capital cities. We can also observe that the northern region has a smaller number of towns per area when compared to the southern and southeastern regions of the country.

Fig. 15 shows a choropleth map of the number of occupational accidents by 1000 occupied people in each Brazilian state, that groups the cities shown in Fig. 14. We can observe that the states with the highest number of accidents per employed population are located in the country's center-south.

The states with the highest GDP in Brazil are São Paulo (SP), Rio de Janeiro (RJ), Minas Gerais (MG), Paraná (PR) and Rio Grande do Sul (RS). The most populous states in the country are also São Paulo, Minas Gerais, and Rio de Janeiro. Since these states have the largest population, a large working staff and concentrate a large part of the country's productive activities, we should expect that the number of work accidents will be significantly higher in them, which explains the distribution seen in Fig 15. On the other hand, the states of the central-west of

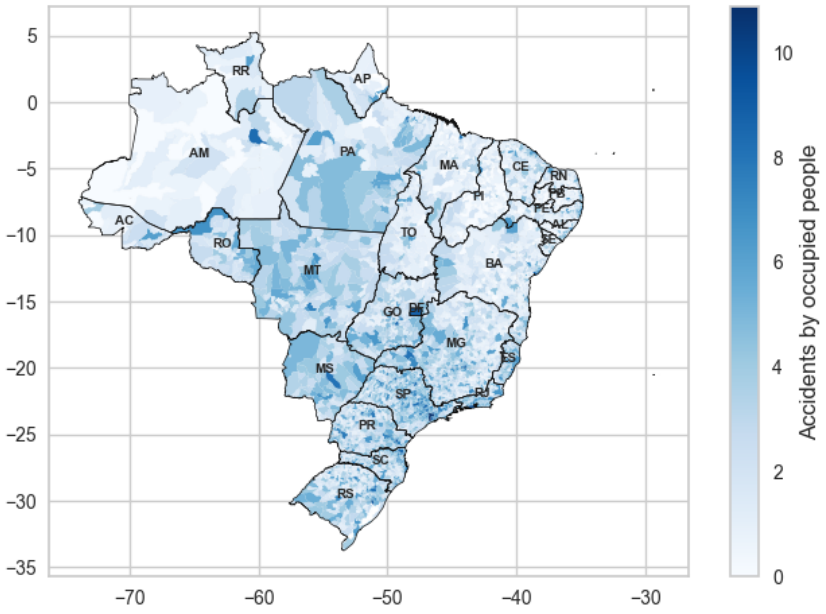


Figure 14 – Occupational accidents by working staff in each Brazilian city.

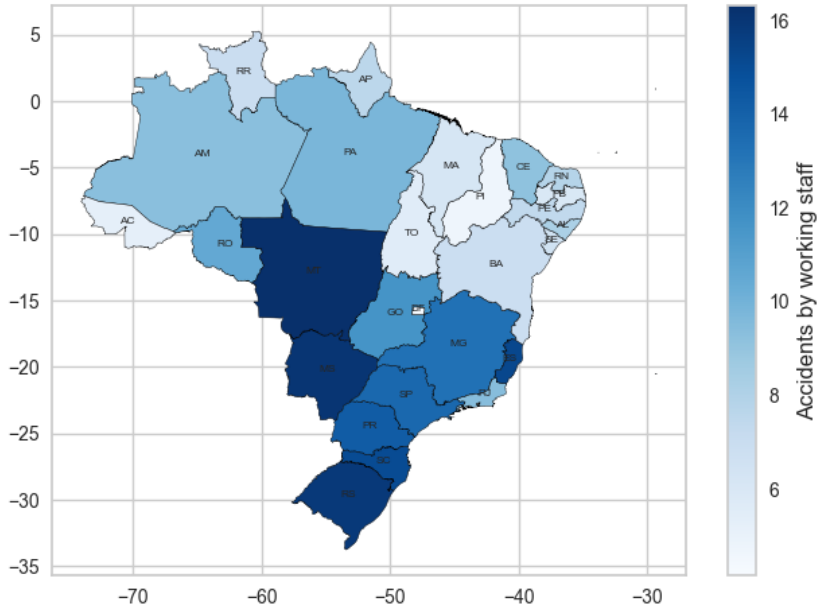


Figure 15 – Accidents by 1000 occupied people.

Brazil, Goiás (GO), Mato Grosso (MT) and Mato Grosso do Sul (MS) are important agricultural producers, developing activities that still require labor and that present risks to the health of workers, which may also indicate the large number of accidents per employed population, as seen in the Fig. 15.

Demographic and economic activity characteristics are important for assessing the statistical distribution of work accidents (TOLEDO; MOURA, 2024) and may be fundamental variables in the proposed work. In Fig. 16, we show bar graphs of the distribution of occupational accidents for the economic activities in which these accidents are the most prevalent in Brazil.

We also show the distribution of accidents by laborer sex. Note that in Brazil, the highest number of occupational accidents is recorded in health services, such as hospitals and dental and medical clinics. This may occur because healthcare professionals are among those responsible for sending information on occupational accidents.

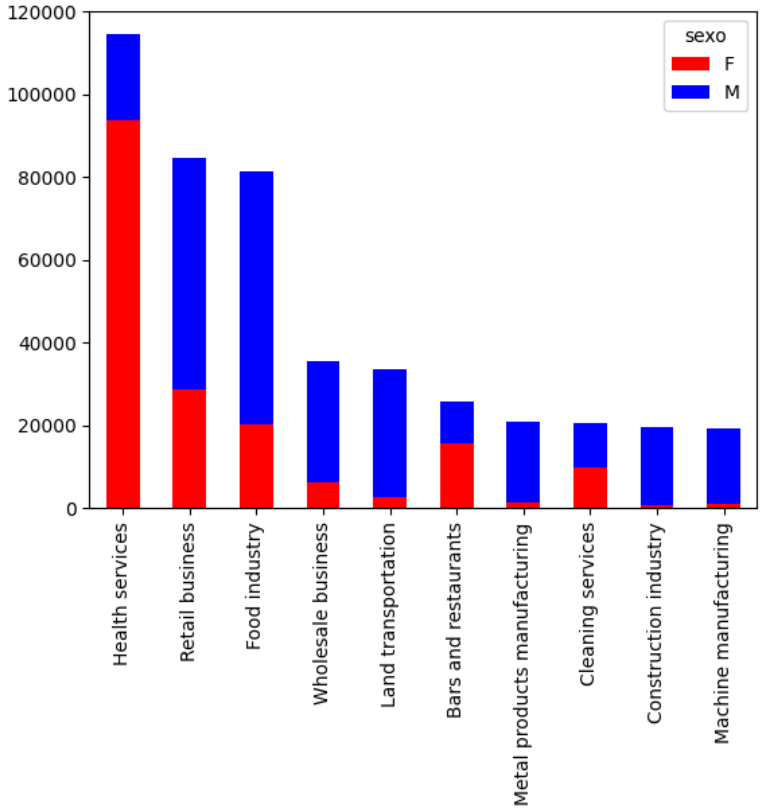
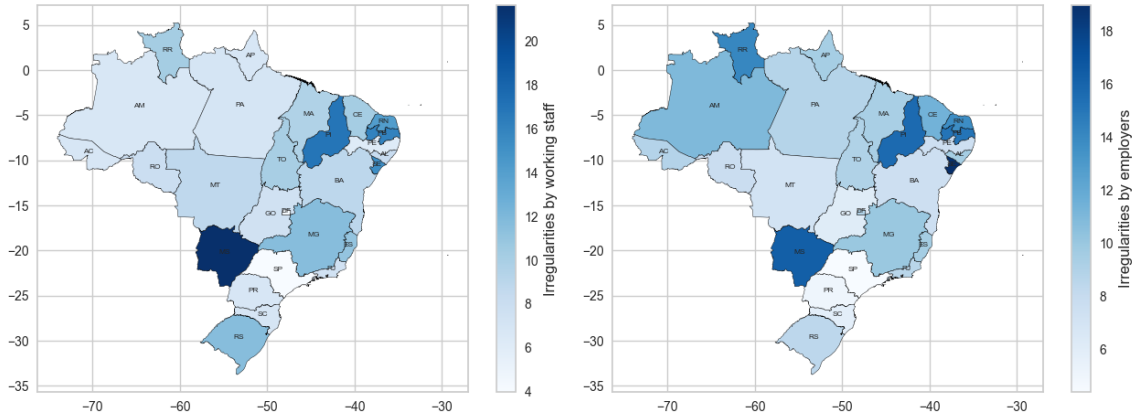


Figure 16 – Accidents by economic activity and sex.

In Brazil, most workplace accidents occur among young men (TOLEDO; MOURA, 2024). In Fig. 16, we can see that the fact that workplace accidents affect more male workers is true in most economic activities - in 8 (eight) out of the 10 (ten) economic activities shown in Fig. 16 the accidents are more prevalent among men. In health services, on the other hand, accidents affect more female workers, who occupy more nursing positions.

Another possible factors that can help predict occupational accidents in Brazil are the labor irregularities already detected in the country. Fig. 17 represents the distribution of irregularities detected by labor inspection in Brazil. We observe in Fig. 17a a choropleth map of irregularities detected by occupied people and a map of irregularities by 100 employers in Fig. 17b. Unlike Fig. 15, the irregularities are not concentrated in the southern and southeastern regions, probably due to the distribution of labor inspection activities throughout the national territory.

In this work, we aim to obtain a machine-learning model to predict the number of occupational accidents in Brazilian cities. Thus, the target variable is obtained from the CAT dataset as discussed in what follows.



(a) Irregularities by 1000 occupied people. (b) Irregularities by 100 employers.

Figure 17 – Irregularities detected by labor inspection in Brazil.

5.3 Data preprocessing and model training

This section describes the data preprocessing and analysis of the ML model training, as shown in Fig. 13.

5.3.1 Data preprocessing

Before we train ML models, an important task still needs to be accomplished, which is to preprocess the data. In this view, all null numerical data are first replaced by zero.

Since we use linear regression as a baseline model, the categorical variables must be transformed into numerical ones. In this work, we deal with a large number of variable categories and, then, adopt the target encoding strategy (MICCI-BARRECA, 2001), proven to be effective in similar problems (PARGENT et al., 2022). In this strategy, each category is encoded based on a shrunk estimate of the average target values for observations belonging to the category.

Finally, the numerical variables were standardized by the common standard scalar strategy. The standard score (z) of a variable value (x) is obtained by subtracting its value from its mean (μ) and dividing the result by its standard deviation (σ):

$$z = \frac{x - \mu}{\sigma}. \tag{12}$$

After the ETL and preprocessing steps, the resulting dataset has 205,663 rows.

5.3.1.1 Train-test split

Dividing the dataset into train and test sets is a common step used in ML. The first one is used to train the algorithm and adjust its parameters. The predictions are, then, evaluated in the test dataset. In the current work, the dataset resulting from the preprocessing step is randomly

divided into a training dataset (with 80% of the data instances, or 164,530) and a test dataset (with 20% of the data instances, or 41,133).

It is important to highlight that, during the training and hyperparameter tuning phase, we use cross-validation, in which the training dataset is iteratively split into ten subsets, nine of which are used to train the model and the rest used to validate the predictions. With this, the best set of hyperparameters is selected, and the algorithm is trained once again with these hyperparameters and with all the training data.

5.3.2 Model training

In this work, we propose using ML algorithms to predict the number of occupational accidents for each economic activity in all Brazilian cities. We use regression models since we predict a continuous real number. More specifically, we use linear regression as a baseline, decision trees, and gradient boosting algorithms, since they have been winning many ML competitions and presenting high prediction performance (BENTÉJAC; CSÖRGŐ; MARTÍNEZ-MUÑOZ, 2021; NIELSEN, 2016). These algorithms are also used in problems similar to the one developed in this work (TOLEDO; MOURA, 2024; LU et al., 2019; NOIA et al., 2020).

As already discussed, we also intend to discover which variables most influence the predictions of the occurrence of occupational accidents in the country. Thus, as shown in Fig. 13, we propose an interactive feature selection based on the variables' importance.

It is known that gradient boosting models calculate a score for each feature, with a higher score representing a larger effect on the prediction. This score can be understood as the feature importance. The LightGBM model, for example, calculates two types of feature importance: split, which is the default method and measures the number of times the feature is used to split data in a model, and gain, which quantifies the improvement in the model's predictions by using a particular feature for splitting [5]. Thus, in the proposed methodology, we first train the LightGBM model with standard hyperparameters with the training dataset to make an initial prediction and determine the feature importance of Fig. 18.

Then, we train the Decision Tree, GradientBossting, LightGBM, XGBoost, and CatBoost algorithms now performing the choice of models' hyperparameters and iteratively increasing the number of eliminated features, from zero to nine. The choice of ML models' hyperparameters can improve the prediction and we display in Table 11 the hyperparameter search space used for the trained models. We use cross-validation with ten folds and Bayesian optimization to perform the hyperparameter search. In this sense, the training dataset is divided into ten folds, nine of which are used for training the model, while the last is used for evaluating the predictions. After the process of iterations, the best hyperparameters are chosen and the whole dataset is used to train the algorithm, which is evaluated with the test dataset.

[5] https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.plot_importance.html

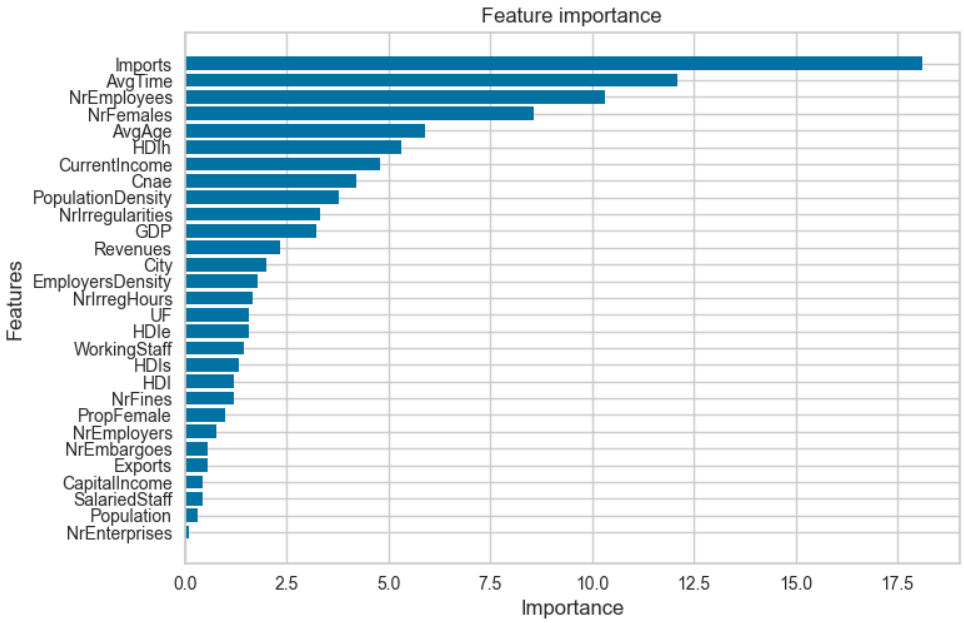


Figure 18 – Feature importance of LightGBM algorithm with standard hyperparameters.

In the proposed methodology, we train the models with hyperparameter tuning using all features of Fig. 18. Thus, we eliminate the least important feature (the number of enterprises), train the models again, perform hyperparameter tuning, and evaluate the predictions. We then remove the second least important variable (population) and train and evaluate the models. We repeat this process until we eliminate 9 (nine) features of Fig. 18.

It is also important to mention that the models trained with hyperparameter tuning are compared with the linear regression algorithm used as a baseline model.

5.4 Results and discussion

This section discusses the results obtained by the ML models trained as shown in Sec. 5.3.

We present the metrics obtained for the models in the test dataset in Tables 12 and 13. Firstly, it is important to mention that the Linear Regression algorithm used as a baseline model presented an RMSE of 86.695 and a R^2 of 0.48350. Thus, we can observe that all other models trained with feature selection and hyperparameter tuning have better predictions than the baseline model since they have higher R^2 and lower values of RMSE.

All models have optimal metrics while removing some features. We can also observe that the LightGBM and Catboost models had similar behavior. As we remove features, the models' prediction improves until they reach an optimal value, from which the metrics worsen. For LightGBM, the best evaluation is obtained when three variables (the number of enterprises, the population, and the salaried staff) are removed, while for the Catboost model, the removal of six variables (the number of enterprises, the population, the salaried staff, the capital income, the

Model	Hyperparameter search space
Decision Tree	max_depth: [5, 9, 21] min_samples_split: [2, 5, 10] min_samples_leaf: [1, 2, 4]
Gradient Boosting	max_depth: [5, 9] max_leaf_nodes: [6, 17] min_samples_leaf: [20, 30] learning_rate: [0.05, 0.5]
XGBoost	max_depth: [5, 9] max_leaves: [6, 17] booster: [gbtree, dart] subsample: [0.7, 0.8, 0.9, 1.0] colsample_bytree: [0.8, 0.9, 1.0] learning_rate: [0.05, 0.5]
LightGBM	max_depth: [5, 9] num_leaves: [6, 17] boosting_type: [gbdt, dart] subsample: [0.7, 0.8, 0.9, 1.0] colsample_bytree: [0.8, 0.9, 1.0] learning_rate: [0.05, 0.5]
Catboost	depth: [5, 9] bagging_temperature: [0.0, 1.0] l2_leaf_reg: [2, 30] border_count: [1, 255] random_strength: [1e-09, 10] learning_rate: [0.05, 0.5]

Table 11 – Hyperparameter search spaces. .

total of exports, and the number of embargoes) presented predictions with better results. The XGBoost algorithm, on the other hand, shows slightly different behavior, with no uniform trend of improving predictions being observed as features were removed. Even so, XGBoost showed better results with the removal of five features.

The high values of R^2 shown in Tabs. 12 and 13 indicate the distancing of the prediction from random guess and, consequently, suggest that the chosen features explain the number of accidents through the algorithms. The increasing value of R^2 while the some variables shown in Fig. 18 are removed, indicates that the feature selection strategy adopted in this work improves the predictions.

Analyzing Tabs. 12 and 13, we can observe only a 1 to 2 % percentual difference between the actual and predicted values of occupational accidents. The values of MAPE were small in most models trained, confirming the robustness of the predictions presented in this paper.

Analyzing the feature selection strategy and its influence on models' performance is also important. As mentioned, removing some features increases the performance of algorithms' predictions, with the optimal number of remaining features dependent on the trained algorithm.

Decision Trees			
Removed features	R^2	MAPE	RMSE
0	0.58205	1.3557	66.676
1	0.63673	1.3384	64.931
2	0.79853	1.3293	47.451
3	0.67198	1.3171	61.491
4	0.42946	1.2533	88.525
5	0.70149	1.2553	51.748
6	0.77990	1.2832	47.184
7	0.51895	3.0908	77.730
8	0.68371	1.3194	53.870
9	0.43683	1.3849	83.751
Gradient Boosting			
Removed features	R^2	MAPE	RMSE
0	0.77187	1.9476	47.316
1	0.77229	1.7704	38.068
2	0.71070	2.1566	73.383
3	0.74205	1.9635	54.020
4	0.77504	2.0831	58.421
5	0.79722	1.8061	51.529
6	0.85043	1.7639	43.554
7	0.73985	1.9538	46.624
8	0.62660	2.1065	53.224
9	0.69880	2.0725	67.163

Table 12 – Models' evaluations

In Fig. 18, we can observe that city population and salaried staff are respectively the second and third least important features and are removed in the proposed strategy, increasing model performance. On the other hand, the working staff is maintained, probably because it is a variable highly correlated to the population and salaried staff. Similar behavior is observed for different groups of variables. While capital income and total exports are removed in most successful models, the total imports and cities' current income are maintained. As to labor inspection variables, removing the number of embargoes and the number of fines also increases the algorithms' performance.

In Table 14, we list the best hyperparameters obtained for the algorithms trained considering the number of excluded features for which the model performs best.

In Fig. 19, we represent the prediction error plots for the trained models in the test dataset. On the horizontal axis of the graphs, we depict the actual values of the target variable, while on the vertical axis, we observe the predicted values for the variable. The identity line in these plots represents a scenario where the prediction matches the model exactly. Observing Fig. 19, we can observe how much variance exists in the predictions. In Fig. 19, we notice that the DT algorithm shows a higher prediction error when compared to gradient boosting algorithms.

LightGBM			
Removed features	R^2	MAPE	RMSE
0	0.83020	2.0490	55.566
1	0.88011	1.9774	47.266
2	0.86172	1.9860	50.516
3	0.89102	1.7096	45.197
4	0.87391	1.9746	48.395
5	0.84212	2.0371	59.783
6	0.75083	1.9042	55.336
7	0.74239	1.8372	56.267
8	0.75238	2.0025	55.164
9	0.75238	2.0025	55.164
XBoost			
Removed features	R^2	MAPE	RMSE
0	0.92862	1.820	34.019
1	0.94733	1.6111	39.029
2	0.83947	1.6467	61.762
3	0.94274	1.4154	32.026
4	0.89044	1.7740	47.903
5	0.94722	1.4099	38.978
6	0.93222	1.3082	34.446
7	0.93432	1.7441	44.673
8	0.92680	1.4247	43.604
9	0.85450	1.4732	45.516
Catboost			
Removed features	R^2	MAPE	RMSE
0	0.66967	1.6515	87.261
1	0.89141	1.7592	36.314
2	0.66243	1.8284	51.318
3	0.87170	1.7389	43.577
4	0.79925	1.7334	49.281
5	0.71289	1.7979	62.765
6	0.90997	1.8031	34.273
7	0.85021	1.6465	46.898
8	0.84285	1.9452	57.478
9	0.83851	1.5235	56.631

Table 13 – Models' evaluations

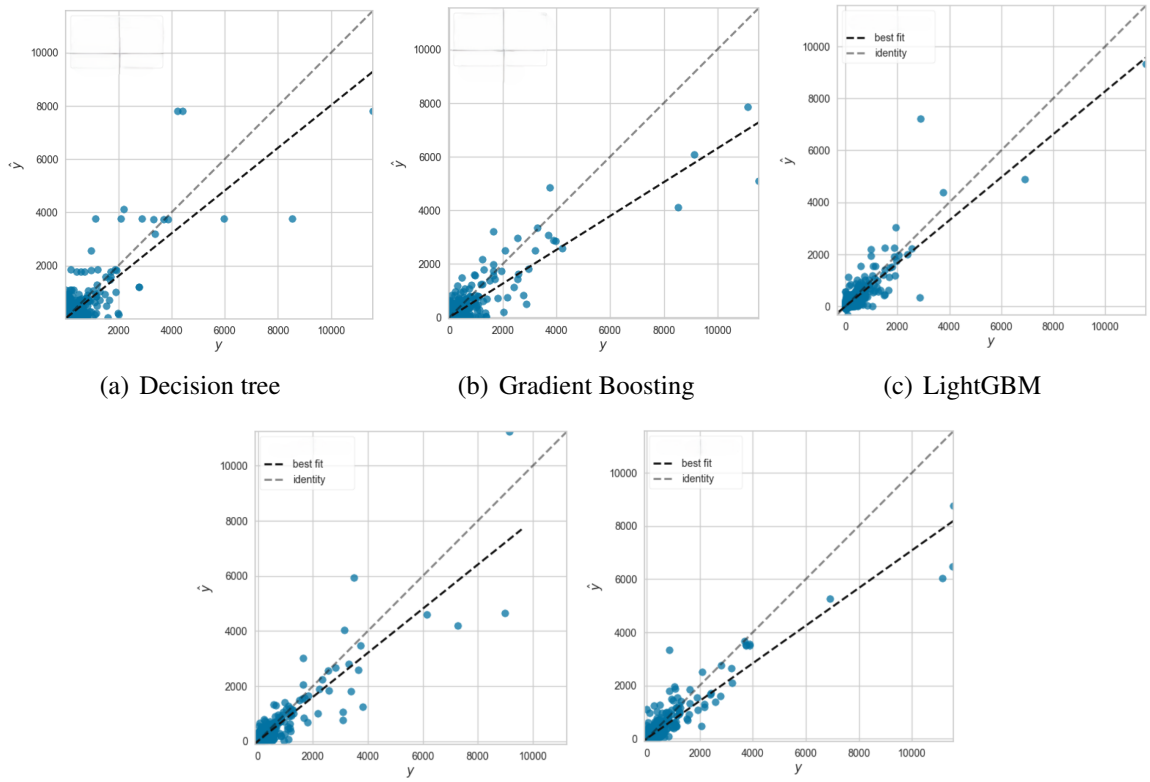
Model	Best hyperparameters
Decision Tree	max_depth: 21 min_samples_split: 10 min_samples_leaf: 4
Gradient Boosting	max_depth: 7 max_leaf_nodes: 6 min_samples_leaf: 21 learning_rate: 0.2469
XGBoost	max_depth: 9 max_leaves: 15 booster: dart subsample: 0.9 colsample_bytree: 1.0 learning_rate: 0.2426
LightGBM	max_depth: 9 num_leaves: 17 boosting_type: dart subsample: 1.0 colsample_bytree: 1.0 learning_rate: 0.3885
Catboost	depth: 7 bagging_temperature: 0.5731 l2_leaf_reg: 30 border_count: 74 random_strength: 1e-09 learning_rate: 0.6196

Table 14 – Best hyperparameters.

Although Table 13 shows differences in models' evaluation metrics, for a fixed number of removed features, these differences may be caused by statistical flukes. Thus, we also use the paired Wilcoxon signed rank test (CONOVER, 1999) to compare the performance of models. This statistical test is the nonparametric version of the Student's t-test and does not assume that the distributions are Gaussian. In this hypothesis test, we assume a null hypothesis (H_0) which states that the ML models have the same performance, while the alternative hypothesis (H_1) states a significant difference between the models. Considering a significance level $\alpha = 0.05$, if the p-value is smaller than α, we reject the null hypothesis and accept that there is a significant difference between the two models (DIETTERICH, 1998).

We perform the tests comparing pairs of models, with a fixed number of excluded features that lead to the best evaluation metrics. For the Wilcoxon test, we use 10-fold cross-validation in the training dataset. The results of the tests are displaced in Tab. 15.

Firstly, we compare the XGBoost model, which presented greater R^2, with the most simple algorithms trained (DT and Gradient Boosting). In both cases, we removed 5 (five) features and the Wilcoxon test showed that the differences between the model's performance



(a) Decision tree (b) Gradient Boosting (c) LightGBM

(d) Gradient Boosting (e) Catboost

Figure 19 – Predction error plots.

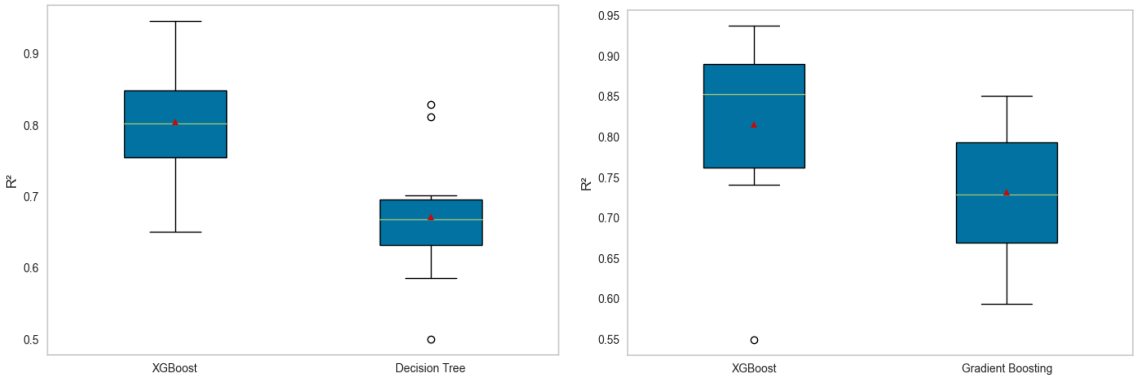
Models	p-value	Interpretation
XGBoost - Decision Trees	0.00195	The models present different performances
XGBoost - Gradient boosting	0.01367	The models present different performances
LightGBM - XGBoost	0.43164	The models present similar performance
LightGBM - Catboost	0.04882	The models present different performances
XGBoost - Catbosst	0.10547	The models present similar performance

Table 15 – Wilcoxon signed-rank test results.

are probably real since the p-value for the XGBoost-DT test was 0.00195 and 0.01367 for the XGBoost-Gradient Boost comparison.

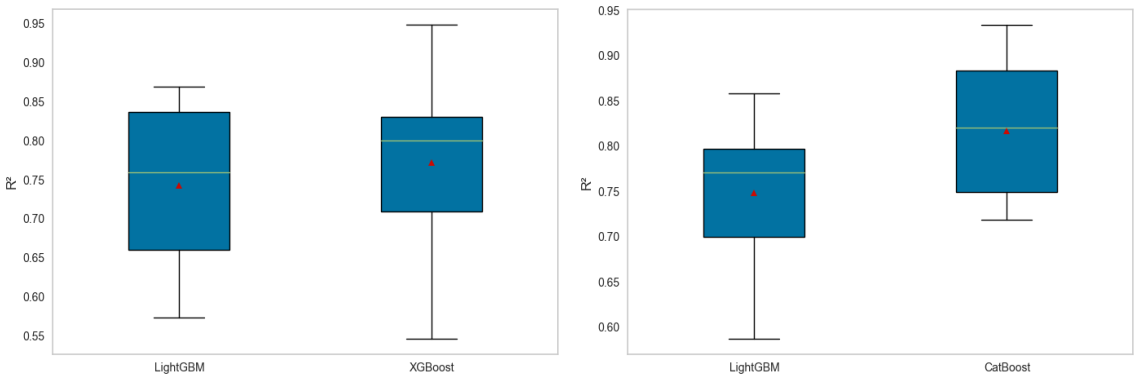
For comparing the algorithms LightGBM and XGBoost, we removed 5 features and used the best hyperparameters of Tab. 14. The Wilcoxon test presents a p-value of 0.43164, greater than the considered α. So, for this pair of algorithms, we must accept the null hypothesis and, thus, the difference between the models' performances is probably not real. On the other hand, comparing the algorithms LightGBM and CatBoost while removing 6 features, we obtain a p-value of 0.04882. Since the p-value $< \alpha$, the null hypothesis must be rejected, and the models' differences are probably real. Finally, comparing XGBoost and Catboost while removing the 6 least important features, we obtain a p-value around 0.10547, which means that the models are probably similar.

In Figs. 20 and 21, we represent the boxplot of R^2 calculated in the folds for the algorithms

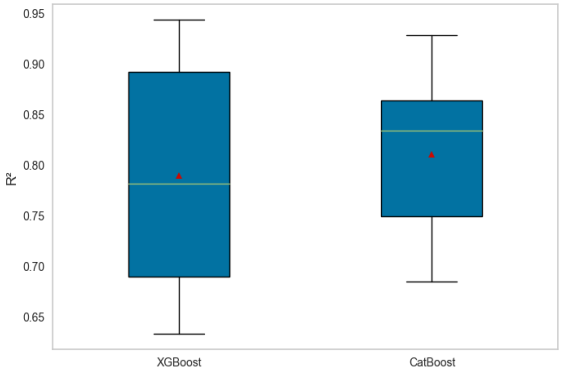


(a) XGBoost-Decision Trees (b) XGBoost-Gradient boosting

Figure 20 – Boxplots of R^2 for tests with the trained algorithms.



(a) LightGBM-XGBoost (b) LightGBM-Catboost



(c) XGBoost-Catboost

Figure 21 – Boxplots of R^2 for tests with the trained algorithms.

compared. In each graph, we plot the boxplots comparing the pairs of algorithms as described in what follows. We can observe in Fig. 21 that CatBoost presents greater uniformity of predictions, with smaller differences between the average value of R^2 and those obtained in training with each folder.

6 CONCLUDING REMARKS

In this chapter, we present the conclusions and the possible future works.

6.1 Conclusions

In Chapter 3, we obtained a single dataset containing socioeconomic variables of the Brazilian cities, coming from multiple sources and covering different areas of knowledge. The problem of non-standardization of columns in the sources was treated in the transform stage of ETL, which allowed us to obtain a dataset with a unique key for each municipality. The final dataset has data related to the economy, population, and public health.

Brazil is one of the largest countries in the world, full of socioeconomic inequalities and geographic diversity in its territory. The core contribution of this work is, then, to represent such diversity in terms of data, simplifying the data mining process for data scientists, government, and companies.

As can be seen, the data update can be very challenging, since it depends on the research institutes that produce and make the variables available. In the same way, the completeness of the dataset depends on these organs. So, in possible future contributions, it will be needed to verify the modifications in the data sources, update the values, and seek new variables.

In Chap. 4, we statistically analyze occupational accident occurrence in Brazil, showing its distribution in the population. We also show the ETL process to obtain an integrated dataset of socioeconomic variables extracted from multiple sources and used as features. We, thus, integrate this feature dataset with the number of occupational accidents in Brazilian states and use ML models to make predictions.

Analyzing the results described in Sec. 4.4, we can state that the built predictive models can anticipate the number of occupational accidents in Brazilian states with high predictability and small expected errors.

We can conclude the trained models with the selected features explain the target variable due to the high R^2 values obtained for the SVM and gradient boosting algorithms (XGBoost and LightGBM). The low percentual difference between the predicted and actual value of the accident number in Brazilian states is observed by the MAPE values in the order of 1.8% to 3.3%.

If we compare the gradient boosting models, we observe that LightGBM has higher R^2 and RMSE, while XGBoost has a slightly high value of MAPE. Observing Fig. 11, we notice that the LightGBM has predictions with less variance when compared with XGBoost. Fig. 12 depicts the features' importance for the LightGBM algorithm, showing that the open-source

socioeconomic statistical variables play an important role in the model prediction.

Chap. 5 presents the ETL process to obtain an integrated dataset with socioeconomic variables extracted from multiple sources and used as features to predict the number of occupational accidents in Brazilian cities. We, then, analyze statistically the occurrence of occupational accidents in Brazil. A major contribution of this work is integrating data from multiple sources, obtaining a table with several variables, from which the subsequent ML models are trained. The data obtained can be used in this work and for constructing future models.

Analyzing the results described in Section 5.4, we can state that the built predictive models can anticipate the number of occupational accidents in Brazilian cities with high predictability and small expected errors. The trained models with the selected features explain the target variable due to the high R^2 values obtained. The low percentual difference between the predicted and actual value of the accident number in Brazilian states is observed by the MAPE values in the order of 1.4% to 2%. The DT and gradient boosting algorithms (GradientBoosting, LightGBM, XGBoost, and CatBoost) present evaluation metrics superior to the baseline model (linear regression).

The feature selection strategy showed that removing some variables improves the prediction of the trained ML models. In addition to evaluating the models using the proposed metrics, we also use the Wilcoxon signed-rank test to compare the algorithms and evaluate if statistical flukes cause the differences between the predictions.

Although they bring serious material and social consequences, occupational accidents can be prevented, according to the literature. (IVASCU; CIOCA, 2019; ALLI, 2008). According to Brazilian legislation, employers are responsible for providing a safe working environment, and the government is responsible for enforcing the law and proposing preventive actions. Thus, predicting the occurrence of occupational accidents can help the government plan inspections and create public policies. Companies can also use the results presented to direct preventive actions in activities that may cause more accidents. Thus, this work can help not only to understand the occurrence of accidents theoretically but also to adopt measures to avoid them.

6.2 Future works

Finally, it is important to highlight that several studies can still address the topic discussed here, expanding the results presented. For example, the temporal behavior of occupational accidents can be evaluated using time series analysis techniques. These topics may be the subject of future contributions.

REFERENCES

ALLI, B. O. *Fudamental Principles of Occupational Health and Safety*. [S.l.: s.n.], 2008. Citado 2 vezes nas páginas 13 e 63.

ALPAYDIN, E. *Machine learning*. [S.l.]: Mit Press, 2021. Citado 3 vezes nas páginas 14, 17 e 18.

BAIG, M. M. et al. Performance comparison of catboost and random forest algorithms for breast cancer prediction: A literature review. In: IEEE. *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*. [S.l.], 2023. p. 1–6. Citado na página 21.

BANDY, G. *International public financial management: Essentials of public sector accounting*. [S.l.]: Routledge, 2018. Citado na página 29.

BENTÉJAC, C.; CSÖRGŐ, A.; MARTÍNEZ-MUÑOZ, G. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, Springer, v. 54, p. 1937–1967, 2021. Citado 3 vezes nas páginas 20, 42 e 54.

BREIMAN, L. *Classification and regression trees*. [S.l.]: Routledge, 2017. Citado na página 19.

BROWNLEE, J. Xgboost with python. *Machine Learning Mastery*, 2019. Citado na página 19.

CHARAPAQUI-MIRANDA, S. et al. Comparing predictive machine learning algorithms in fit for work occupational health assessments. In: SPRINGER. *Annual International Symposium on Information Management and Big Data*. [S.l.], 2019. p. 218–225. Citado 2 vezes nas páginas 22 e 26.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794. Citado 2 vezes nas páginas 19 e 20.

CHENG, M.-Y.; KUSOEMO, D.; GOSNO, R. A. Text mining-based construction site accident classification using hybrid supervised machine learning. *Automation in Construction*, Elsevier, v. 118, p. 103265, 2020. Citado 2 vezes nas páginas 24 e 26.

CONOVER, W. J. *Practical nonparametric statistics*. [S.l.]: john wiley & sons, 1999. v. 350. Citado na página 59.

DIETTERICH, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 10, n. 7, p. 1895–1923, 1998. Citado na página 59.

DOROGUSH, A. V.; ERSHOV, V.; GULIN, A. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018. Citado na página 21.

FEURER, M.; HUTTER, F. Hyperparameter optimization. In: *Automated machine learning*. [S.l.]: Springer, Cham, 2019. p. 3–33. Citado na página 21.

FISCHER, T. K. et al. A mortalidade infantil no brasil: série histórica entre 1994-2004 e associação com indicadores socioeconômicos em municípios de médio e grande porte. *Medicina (Ribeirão Preto)*, v. 40, n. 4, p. 559–566, 2007. Citado 2 vezes nas páginas 27 e 29.

FREUND, Y.; SCHAPIRE, R.; ABE, N. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, JAPANESE SOC ARTIFICIAL INTELL, v. 14, n. 771-780, p. 1612, 1999. Citado na página 19.

GANGULI, R.; MILLER, P.; POTHINA, R. Effectiveness of natural language processing based machine learning in analyzing incident narratives at a mine. *Minerals*, MDPI, v. 11, n. 7, p. 776, 2021. Citado 2 vezes nas páginas 24 e 26.

GHOLAMIZADEH, K. et al. An integration of intelligent approaches and economic criteria for predictive analytics of occupational accidents. *Decision Analytics Journal*, Elsevier, v. 9, p. 100357, 2023. Citado na página 26.

IBGE. *Brasil | Cidades e Estados - IBGE*. 2023. <https://www.ibge.gov.br/cidades-e-estados>. Accessed: 2023-06-15. Citado na página 14.

IBGE. *Conheça o Brasil - biomas brasileiros*. 2023. <https://educa.ibge.gov.br/jovens/conheca-o-brasil/territorio/18307-biomas-rasileiros.html>. Accessed: 2023-06-15. Citado na página 14.

IBGE. *Conheça o Brasil - clima*. 2023. <https://educa.ibge.gov.br/jovens/conheca-o-brasil/territorio/20644-clima.html>. Accessed: 2023-06-15. Citado na página 14.

IBGE. *IBGE - Áreas Territoriais*. 2023. <https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15761-areas-dos-municipios.html?=&t=o-que-e>. Accessed: 2023-06-15. Citado na página 14.

IBGE. *Instituto Brasileiro de Geografia e Estatística - IBGE*. 2023. <https://www.ibge.gov.br/>. Accessed: 2023-06-15. Citado na página 27.

IBRAHIM, A. A. et al. Comparison of the catboost classifier with other machine learning methods. *International Journal of Advanced Computer Science and Applications*, Science and Information (SAI) Organization Limited, v. 11, n. 11, 2020. Citado na página 21.

IPEA. *Instituto de Pesquisa Econômica Aplicada - IPEA*. 2023. <https://www.ipea.gov.br/portal/>. Accessed: 2023-06-15. Citado na página 27.

IVASCU, L.; CIOCA, L.-I. Occupational accidents assessment by field of activity and investigation model for prevention and control. *Safety*, MDPI, v. 5, n. 1, p. 12, 2019. Citado 2 vezes nas páginas 13 e 63.

JAEN-VARAS, D. et al. The association between adolescent suicide rates and socioeconomic indicators in brazil: a 10-year retrospective ecological study. *Brazilian Journal of Psychiatry*, SciELO Brasil, v. 41, p. 389–395, 2019. Citado na página 27.

JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112. Citado 6 vezes nas páginas 17, 18, 19, 21, 22 e 42.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015. Citado na página 17.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, v. 30, 2017. Citado 2 vezes nas páginas 20 e 21.

KHAIRUDDIN, M. Z. F. et al. Occupational injury risk mitigation: machine learning approach and feature optimization for smart workplace surveillance. *International journal of environmental research and public health*, MDPI, v. 19, n. 21, p. 13962, 2022. Citado 2 vezes nas páginas 23 e 26.

KOC, K.; EKMEKCIOĞLU, Ö.; GURGUN, A. P. Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers. *Automation in Construction*, Elsevier, v. 131, p. 103896, 2021. Citado 2 vezes nas páginas 23 e 26.

KOKLONIS, K. et al. Utilization of machine learning in supporting occupational safety and health decisions in hospital workplace. *Engineering, Technology & Applied Science Research*, v. 11, n. 3, p. 7262–7272, 2021. Citado 2 vezes nas páginas 24 e 26.

LAAT, P. B. D. Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? *Philosophy & technology*, Springer, v. 31, n. 4, p. 525–541, 2018. Citado na página 13.

LIASHCHYNSKYI, P.; LIASHCHYNSKYI, P. Grid search, random search, genetic algorithm: A big comparison for nas. *arXiv preprint arXiv:1912.06059*, 2019. Citado na página 21.

LU, Y. et al. A comparative study on the prediction of occupational diseases in china with hybrid algorithm combing models. *Computational and mathematical methods in medicine*, Hindawi Limited, v. 2019, 2019. Citado 3 vezes nas páginas 24, 26 e 54.

MACIEJEWSKI, M. To do more, better, faster and more cheaply: Using big data in public administration. *International Review of Administrative Sciences*, SAGE Publications Sage UK: London, England, v. 83, n. 1_suppl, p. 120–135, 2017. Citado na página 13.

MAMMONE, A.; TURCHI, M.; CRISTIANINI, N. Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 1, n. 3, p. 283–289, 2009. Citado na página 19.

MERGEL, I.; RETHEMEYER, R. K.; ISETT, K. Big data in public affairs. *Public Administration Review*, Wiley Online Library, v. 76, n. 6, p. 928–937, 2016. Citado na página 13.

MICCI-BARRECA, D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, ACM New York, NY, USA, v. 3, n. 1, p. 27–32, 2001. Citado 2 vezes nas páginas 41 e 53.

MITCHELL, T. M. *Machine learning*. 1997. Citado na página 17.

MPT. *Observatório de Segurança e Saúde no Trabalho*. 2023. Accessed: 2023-10-02. Disponível em: <https://smartlabbr.org/sst>. Citado na página 13.

NIELSEN, D. *Tree boosting with xgboost-why does xgboost win"every"machine learning competition?* Dissertação (Mestrado) — NTNU, 2016. Citado 3 vezes nas páginas 20, 42 e 54.

NOBRE, J.; NEVES, R. F. Combining principal component analysis, discrete wavelet transform and xgboost to trade in the financial markets. *Expert Systems with Applications*, Elsevier, v. 125, p. 181–194, 2019. Citado na página 20.

NOIA, A. D. et al. Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction. *Soft Computing*, Springer, v. 24, n. 6, p. 4393–4406, 2020. Citado 4 vezes nas páginas 24, 26, 42 e 54.

ORGANIZATION, W. H. et al. Who/ilo joint estimates of the work-related burden of disease and injury, 2000–2016: global monitoring report. World Health Organization, 2021. Citado 3 vezes nas páginas 13, 39 e 49.

PARGENT, F. et al. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Computational Statistics*, Springer, v. 37, n. 5, p. 2671–2692, 2022. Citado 2 vezes nas páginas 41 e 53.

PROKHORENKOVA, L. et al. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, v. 31, 2018. Citado na página 21.

RECAL, F.; DEMIREL, T. Comparison of machine learning methods in predicting binary and multi-class occupational accident severity. *Journal of Intelligent & Fuzzy Systems*, IOS Press, v. 40, n. 6, p. 10981–10998, 2021. Citado 2 vezes nas páginas 23 e 26.

RODRÍGUEZ-RUEDA, P. et al. Origin–destination matrix estimation and prediction from socioeconomic variables using automatic feature selection procedure-based machine learning model. *Journal of Urban Planning and Development*, American Society of Civil Engineers, v. 147, n. 4, p. 04021056, 2021. Citado na página 27.

ROSSUM, G. V.; JR, F. L. D. *Python tutorial*. [S.l.]: Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995. v. 620. Citado na página 29.

ROSSUM, G. V. et al. Python programming language. In: SANTA CLARA, CA. *USENIX annual technical conference*. [S.l.], 2007. v. 41, n. 1, p. 1–36. Citado na página 38.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, IBM, v. 3, n. 3, p. 210–229, 1959. Citado na página 17.

SANTOS, E. G. d. O.; BARBOSA, I. R. Conglomerados espaciais da mortalidade por suicídio no nordeste do brasil e sua relação com indicadores socioeconômicos. *Cadernos Saúde Coletiva*, SciELO Brasil, v. 25, p. 371–378, 2017. Citado na página 27.

SARKAR, S. et al. Application of optimized machine learning techniques for prediction of occupational accidents. *Computers & Operations Research*, Elsevier, v. 106, p. 210–224, 2019. Citado 2 vezes nas páginas 23 e 26.

SAU, A.; BHAKTA, I. Screening of anxiety and depression among seafarers using machine learning technology. *Informatics in Medicine Unlocked*, Elsevier, v. 16, p. 100228, 2019. Citado 2 vezes nas páginas 25 e 26.

SAúDE, M. da. *DATASUS*. 2023. <https://datasus.saude.gov.br/>. Accessed: 2023-06-15. Citado na página 27.

SCHAPIRE, R. E. et al. A brief introduction to boosting. In: CITESEER. *Ijcai*. [S.l.], 1999. v. 99, n. 999, p. 1401–1406. Citado 2 vezes nas páginas 19 e 20.

SCOTT, E. et al. The development of a machine learning algorithm to identify occupational injuries in agriculture using pre-hospital care reports. *Health information science and systems*, Springer, v. 9, p. 1–9, 2021. Citado 3 vezes nas páginas 23, 24 e 26.

SETHI, I. et al. Machine learning in government applications: A review. *Procedia Computer Science*, Elsevier, v. 258, p. 1365–1371, 2025. Citado na página 13.

TANG, W. et al. Machine learning approach to uncovering residential energy consumption patterns based on socioeconomic and smart meter data. *Energy*, Elsevier, v. 240, p. 122500, 2022. Citado na página 27.

TOLEDO, J.; MOURA, T. J.; TIMOTEO, R. Brstats: a socioeconomic statistics dataset of the brazilian cities. In: SBC. *Anais do V Dataset Showcase Workshop*. [S.l.], 2023. p. 67–78. Citado 4 vezes nas páginas 16, 39, 48 e 49.

TOLEDO, J.; TIMOTEO, R. D. A.; BARBOSA, E. S. Inteligência artificial para predição de acidentes de trabalho no brasil e sua aplicação pela inspeção do trabalho. *Revista da Escola Nacional da Inspeção do Trabalho*, 2020. Citado 3 vezes nas páginas 24, 26 e 42.

TOLEDO, J. M.; MOURA, T. J. M. Occupational accidents prediction in brazilian states: A machine learning based approach. In: *Proceedings of the 26th International Conference on Enterprise Information Systems*. [S.l.: s.n.], 2024. v. 1, p. 595–602. Citado 15 vezes nas páginas 7, 8, 16, 36, 37, 38, 40, 42, 43, 44, 45, 48, 51, 52 e 54.

TOLEDO, J. M.; MOURA, T. J. M. Occupational accidents prediction in brazil: an approach using regression machine learning algorithms. *Lecture Notes in Business Information Processing (accepted paper)*, 2025. Citado na página 16.

TOLEDO, J. M.; MOURA, T. J. M. Predicting occupational accidents in brazilian cities: a machine learning based approach. *International Journal of Data Science and Analytics*, 2025. Under submission. Citado na página 16.

TURNER, R. et al. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In: PMLR. *NeurIPS 2020 Competition and Demonstration Track*. [S.l.], 2021. p. 3–26. Citado na página 21.

WANG, P.; ZONG, L. Does machine learning help private sectors to alarm crises? evidence from china's currency market. *Physica A: Statistical Mechanics and its Applications*, Elsevier, p. 128470, 2023. Citado na página 13.

WU, J. et al. Hyperparameter optimization for machine learning models based on bayesian optimization. *Journal of Electronic Science and Technology*, Elsevier, v. 17, n. 1, p. 26–40, 2019. Citado na página 21.

ZHANG, C. et al. Machine learning based prediction for china's municipal solid waste under the shared socioeconomic pathways. *Journal of Environmental Management*, Elsevier, v. 312, p. 114918, 2022. Citado na página 27.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
PARAÍBA
Campus João Pessoa

PPGTI
Programa de Pós Graduação em Tecnologia da Informação

DECLARAÇÃO

DECLARO, para os devidos fins, que foram realizadas as considerações sugeridas pela Banca Examinadora formada por mim, pela Profa. Dra. Damires Yluska de Souza Fernandes, Examinadora Interna, e pelo Prof. Dr. Yuri de Almeida Malheiros Barbosa, Examinador Externo, após a apresentação do trabalho intitulado “Predicting occupational accidents in Brazil: a machine learning based approach”, requisito parcial para obtenção de título de Mestre em Tecnologia da Informação ao discente Jefferson de Morais Toledo.

João Pessoa, 04 de julho de 2025.

gov.br
Documento assinado digitalmente
THIAGO JOSE MARQUES MOURA
Data: 25/08/2025 16:50:36-0300
Verifique em https://validar.iti.gov.br

Thiago José Marques Moura

Orientador(a) e Presidente da Banca Examinadora