ISSN 1980-5411 (On-line version)
Production, 33, e20220117, 2023
DOI: 10.1590/0103-6513.20220117

Research Article

The automotive recall data search and its analysis applying machine learning

Bruno Fernandes Maione[a*], Paulo Carlos Kaminski[a], Emilio Carlos Baraldi[a]
[a]Universidade de São Paulo, São Paulo, SP, Brasil
*bruno.maione@usp.br

Abstract

Paper aims: This article investigates the worldwide trend of growth in the number of recalls, as well as in the number of products involved in each campaign.

Originality: To investigate these facts, a study of the automotive recall was developed, comprising Brazil, the European Union, and the United States of America.

Research method: Due to the different availabilities between the locations, search tools and software were developed to obtain and group hidden data from 2010 to 2019.

Main findings: In this work, the impacts of the recall were analyzed using three categories of algorithms: clustering, classification, and regression. Analyzes were made about the results obtained and discussions were built about the importance of applying the machine learning technique.

Implications for theory and practice: The use of search tools and software to obtain and group hidden data in databases and opens the opportunity for new research in various areas.

Keywords

Recall. Automotive Industry. Machine Learning. Product Quality.

How to cite this article: Maione, B. F., Kaminski, P. C., & Baraldi, E. C. (2023). The automotive recall data search and its analysis applying machine learning. *Production*, *33*, e20220117. https://doi.org/10.1590/0103-6513.20220117

Received: Nov. 26, 2022; Accepted: Apr. 21, 2023.

1. Introduction

It is noticeable the community's growing interest in failures in the quality of products and services worldwide, especially in serious defects, which may endanger people's physical integrity. Following this same trend, the academic area shows consistent growth in the number of works related to recalls. Such focus direction can be seen in Kalaignanam et al. (2013) work, which tested, for the automobile industry market in the United States of America, the direct effect of recall on future accidents and the future indirect effect on product reliability for the years 1985 to 2013. Subsequently, continuing the work, Eilert et al. (2017) studied the effect of the severity of the problem associated with the recall in the same country, and the time required for its start, considering the years 2000 to 2017.

In principle, it can be said that a recall is carried out when it is discovered or there is suspicion those defects may impair the performance and/or the safety of products or services. As a rule, a recall must reach all its consumers, including those who have not yet had any problems associated with the defect (Eilert et al., 2017). Thus, it is necessary to define what is product safety. A safe product is defined as a product that, when used under normal or reasonably foreseeable conditions (including duration), does not offer any risk of injury, damage to the health of users, property, or the environment (European Union, 2002; Zhu et al., 2018).

Based on the definitions exposed for Baraldi & Kaminski (2016), Eilert et al. (2017), Hora et al. (2011), and Mackelprang et al. (2015), recalls can be divided into two main causes, namely: defects arising from failures in product development and/or the manufacturing process.

In terms of damages, for Eilert et al. (2017) and Hora et al. (2011), recalls imply social costs, such as damage to property and person. Besides, the authors cite that the increase in the number of recalls and the delay of manufacturers in their implementation are attracting the attention of society, and they ask why it takes so long to discover that a product poses a risk to people's safety.

In financial terms, other important damage from the occurrence of recalls, it can be said that the composition of the costs of a recall goes far beyond the costs of correcting a product's non-conformity (Kumar & Schmitz, 2011; Slack et al., 2010). In this case, other factors must be considered, such as the association of the brand with quality problems, customer dissatisfaction with having purchased a defective product, customer discomfort in taking the product to be repaired, a possible risk to people's safety, the possibility of a loss of the market, legal punishments, and civil liability, among others (Baraldi and Kaminski, 2018). In the automobile industry, several companies have suffered huge financial losses due to serious quality problems related to their products (including the realization of the loss of human lives), culminating in the closing of companies responsible for the facts and other social problems, such as unemployment of their employees (Committee on Commerce, Science, and Transportation, 2019; Conner & Wanasika, 2018; Janssen et al., 2015; Maiorescu, 2016). As an example, the United States Department of Transportation's National Highway Traffic Safety Administration (NHTSA), in December 2012, fined Toyota Motor Corporation US$ 17.35 million for failing to report to the US Federal Government a safety defect in its products promptly (National Highway Traffic Safety Administration, 2018a). Financial costs may also appear differently than usual, the Consumer Product Safety Commission, for instance, estimates that deaths, injuries, and property damage resulting from consumer product incidents cost the United States more than US$1 trillion a year (Consumer Product Safety Commission, 2018).

The problem intensifies as it gains recurrence that did not exist before. Bates et al. (2007) carried out a study on patterns and trends in motor vehicle safety recalls in the United Kingdom. For this, they used a data set based on 23.1 million vehicles registered between 1992 and 2002. According to the authors, 10.8 million vehicles were recalled, which represented 47% of all vehicle registrations in the United Kingdom comprised in the period.

Thus, coupled with an increase in the relevance of product safety, the growing number of recalls may, for Haefele & Westkamper (2014), indicate that products have become more insecure. It is stated, however, that such an indicator may be related to many other factors such as the occurrence of more demanding legal requirements, activities of the authorities that are increasingly more stringent, and different legal requirements in different countries, as discussed in the work of Maione et al. (2021b). Also, global motor vehicle manufacturers are urged to deal with an increasing variety of parts and greater manufacturing complexity in addition to internal and external interfaces in a global production network (Haefele & Westkamper, 2014).

Therefore, with the increase in the numbers of recalls and the rise of the manufacturing complexity process, variables not seen before are becoming relevant to the market of automotive products: most of them are hard to be detected by humans analyses. The interpretation of these variables, using standardized statistical methods, however, results in conclusions known to the automobile industry, such as discussed by Gruber et al. (2021). Still, with the innovation and growth speed of artificial intelligence, allied to Industry 4.0, new possibilities for analysis from the use of a part of it, machine learning, appear and enable new tools for making conclusions, such as shown and reviewed by Silva et al. (2020). A reinforcing factor that leads the academic attention to this accelerated field of studies is the possibility of applying it to several distinguished problems, such as recently discussed by Salazar-Reyna et al. (2022), in the healthcare engineering field, and by Choi et al. (2022) with deep neural network models applied to accountability and prediction of abnormal audit fees.

Machine learning, in a concise way, is a technique based on statistics, since it applies optimized mathematical models of inference from samples to obtain estimates (Alppaydin, 2010). Among the useful tools, based on machine learning techniques, are clustering, classification, and regression.

Clustering is the method of allocating examples in clusters (groups), which normally represent some mechanisms existing in the real-world process, which generated these examples, making them more like each other within some organizations (Baranauskas, 2011).

The second strategy, that is, the classification method, represents, for Kotsiantis et al. (2006), algorithms that predict a discrete characteristic of a given object and, thus, infer labels on it, being the object, in this case, an automotive recall.

The third and final category of algorithms to be used is regression. Algorithms of this niche make the machine estimate values, that is, continuous characteristics, from certain inputs and are, for Wakefield (2020), extremely useful for predictions. For automotive recalls, values such as the number of affected vehicles in a year and the number of recalls, also in a specific year, are possible to be found.

The applicability of Machine Learning shows itself as a possible reliable way of trying to achieve competitive advantages in the automotive recall field. Such a statement can be reassured by supporting analytical capabilities and data visualization characteristics directly related to ML outputs, as discussed by Medeiros & Maçada (2022), in which it was proven that managers can establish policies and strategies to extract value from data and leverage business agility and competitiveness through the use of business analytics and big data visualization.

The objective of this article is to investigate, using basic and efficient machine learning approaches, the worldwide trend of growth in the number of recalls, as well as the growth in the number of products involved in each campaign, and obtain a broader and more assertive view of the problem. For this, a study on automotive recall was developed, covering Brazil, the European Union and the United States of America, in the years 2010 to 2019.

2. Automotive recall at the three locations

There are situations where recalls can have their defects investigated and, in these circumstances, they are supervised by a government institution that, among other activities, enforces regulations and monitors them until their conclusion (Eilert et al., 2017). For a better understanding of these facts, it is important to review the legislation and official data sources in the three regions studied.

The Brazilian Ministry of Justice, the first region of interest, has the mission of guaranteeing and promoting citizenship, justice, and public security, through joint action between the State and society (Brasil, 2018a).

In Brazil, the Ministry of Justice makes available in its database of recall alerts, 18 sectors of activities that must be elected during a recall notification, (Brasil, 2017, 2018b).

In contrast to the Brazilian case, the European Union has a different approach, as it is formed by several nations. The European Commission defines the bloc's general jurisdictions on product safety in Directive 2001/95/EC of the European Parliament and of the Council of 3 December 2001 on general product safety (European Commission, 2018a).

In terms of data, the European Union, the RAPEX system provides a report that contains up to 23 information data for each of the recalls inserted in it, but unfortunately, it does not contain the number of products affected in each of the recalls (European Commission, 2018b).

Moving to the last location, it is known that the USA regulation started with an effort to increase the safety of motor vehicles, on September 9, 1966, the Congress passed the National Motor Vehicle and Traffic Safety Act of 1966 (15 U.S.C. 1381) (Rupp & Taylor, 2002).

Currently, the United States of America (USA) organization uses several federal agencies to alert citizens of dangerous or defective products. To facilitate public access to information, six federal agencies with different jurisdictions came together and created the website <www.recalls.gov> (Recalls, 2018; National Highway Traffic Safety Administration, 2018b).

For a better understanding of how recall management occurs in the three regions, Table 1 was developed, which presents a comparison of the three locations in terms of regulation.

Table 1. Comparison between the three regions.

Region	Brazil	European Union	United States of America
Legislation	Constituição Federal Brasileira (Brazilian Federal Constitution)	European Parliament and the Council of the European Union	Federal Regulations of the United States of America
Law	Artigo 5º, inciso XXXII, da Constituição Federal Brasileira (Article 5, Item XXXII, of the Brazilian Federal Constitution)	Directive 2001/95/EC	National Traffic and Motor Vehicle Safety Act of 1966 (15 U.S.C 1381)
Government agency	Ministério da Justiça (Justice Ministry)	European Commission	National Highway Traffic Safety Administration (NHTSA)
Other government agencies	Secretaria Nacional do Consumidor (National Consumer Secretariat)	National authorities of the members of the European Commission	United States Environmental Protection Agency (EPA)
Other government agencies	Grupo de Estudos Permanentes de Acidentes de Consumo (Permanent Study Group on Consumer Accidents)		Office of Vehicle Safety Research and the Office of Behavioral Safety Research
Access to recall information	Ministério da Justiça (Justice Ministry) – (Secretaria Nacional do Consumidor, 2020b)	Rapid Alert System for dangerous non-food products (RAPEX) - (European Commission, 2020)	U.S. Government recalls - (Recalls, 2018)

Source: Baraldi & Kaminski (2019).

3. Database

With the growing importance of using data in obtaining reports and formulating analyzes, contemporary engineering has brought a new problem for technological advancement and, consequently, for machine learning: the reliability of databases. Like an empirically tested mechanical system, there are important aspects in the construction of artificial intelligence models to be considered in the study of the data analyzed, including the presence of noise and outliers, which can compromise the direction of considerations made, as well as the conclusions reached. More generally, the growing importance of having solid, complete, and clean bases, as well as the obsession in their searches, not only allows them to continue their work but is also responsible for innovating and giving rise to new functionalities, within the world of statistics and machine learning, since its principles (Imielinski & Mannila, 1996).

In technical terms, it is seen that the amount of data related to recalls is not considerably numerous, in terms of entries. Thus, for the present study, they were stored in databases on the Microsoft Access platform and, for intuitive consumption, in Microsoft Excel.

For the Federative Republic of Brazil, the use of the databases provided by the National Consumer Secretariat (Secretaria Nacional do Consumidor, 2020a) bases would not be enough. Thus, trying to obtain more data from the national agency, the Web-Scraping technique was applied, a form of mining that allows the extraction of data from websites, converting them into structured information for later analysis, with a script in Python language on the Brazilian entity's information portal. The portal offers the user the possibility to access the data of past recalls in a sequential and unitary way (Secretaria Nacional do Consumidor, 2020b). The purpose of the program (Maione et al., 2021a), therefore, was to automate the operational work of reading all the platform's entries and transcribing them to a local base, allowing time savings and avoiding human errors.

For the next region, the European Union, obtaining the base was also simple and straightforward. Through the alert search tool on the RAPEX platform, Rapid Alert System for dangerous non-food products, made available by the European institute, it was possible to extract the base to be used in the construction of the models (European Commission, 2020).

Unlike the Brazilian case, the European Union does not provide the number of affected products, which makes it difficult to build more comprehensive models. Thus, the construction of a Python script to obtain this information, as done for the Brazilian database, was not effective since it is not even exposed on the platform's websites. The different ways of data distribution, if compared to the Brazilian and the USA cases, can indicate important causal variables to be studied, such as the laws of the regions, about the different approaches to data opening (Maione et al., 2021b).

In contrast to the scenarios in other regions, the United States of America, through the NHTSA's Office of Defects Investigation (ODI) made available an official database for model building, running from 1966 to 2020 (ODI, 2021).

For a better overview of the data obtained after the search with the aid of Web-Scraping techniques, Table 2 was developed, considering a match of the fields of each database to the list of desired fields. It indicates the present and missing fields of three bases of interest.

Table 2. Comparison between the three databases.

+	Brazil	European Union	United States of America
Recall Title	Present	Present	Present
Identification	Present	Present	Present
Risk Description	Present	Present	Present
Type of Product	Present	Present	Present
Manufacturer	Present	Present	Present
Country of Origin	Present	Present	Absent
Country of Export	Present	Present	Absent
Model	Present	Present	Present
Year of Manufacturer	Present	Present	Present
Number of Affected	Present	Absent	Present
Link to source	Present	Present	Absent
Date of Recall	Present	Present	Present
Type of Alert	Absent	Present	Present
Origin of Alert	Not Valid	Present	Not Valid
Measures taken	Absent	Present	Present
Affected Component	Present	Present	Present

Source: Own elaboration.

Considering the variety and quantity of variables in the datasets, a prior selection process was necessary before applying them in modeling strategies. As priority, the models were built using quantitative variables that were available in all three regions and that fit in the concept of the strategy itself.

In addition, to demonstrate the dimensions of the recall problem, the number of licensed new motor vehicles compared to number of motor vehicles involved in recall for the years 2011 to 2019 are presented in Figure 1.

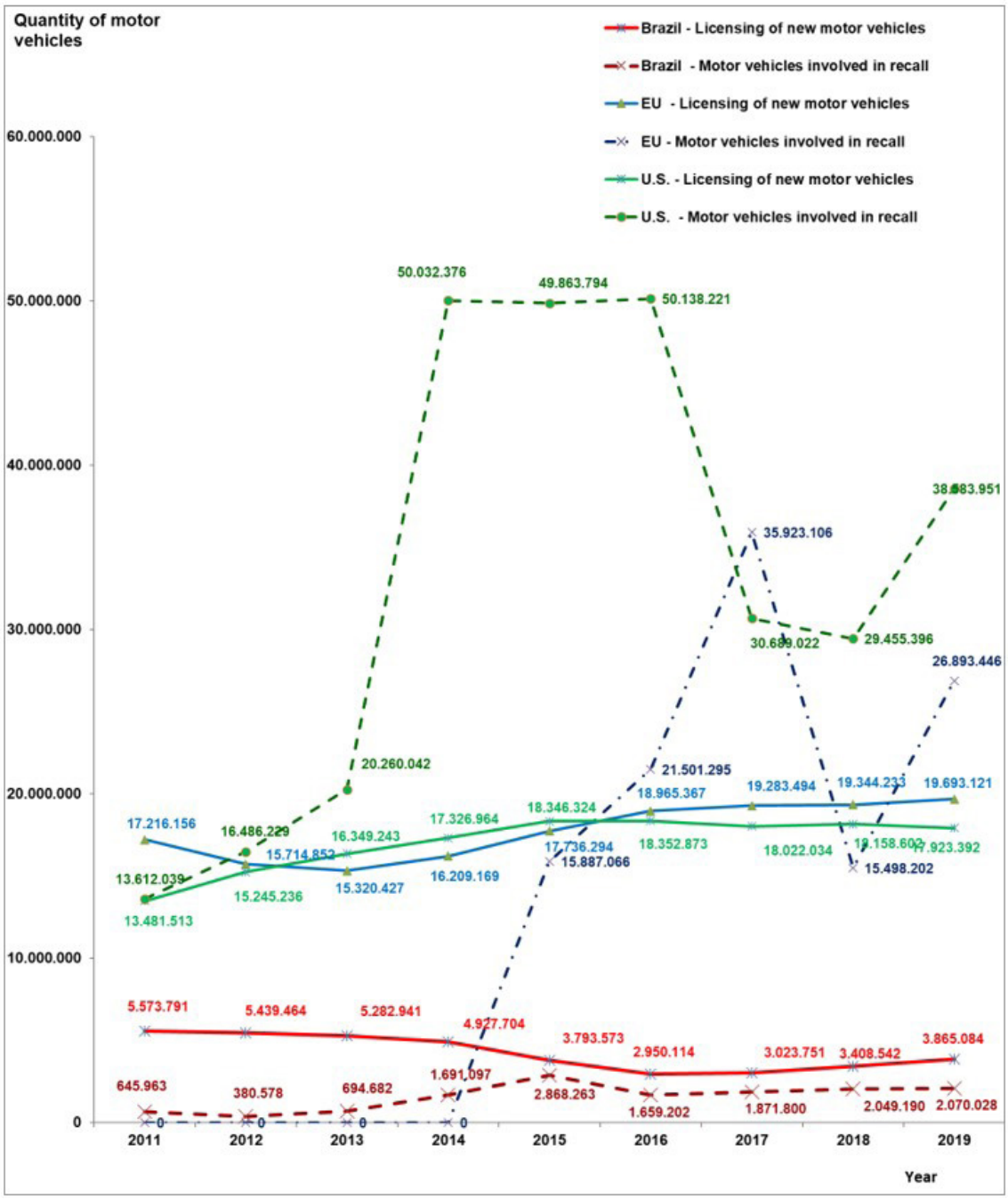

Figure 1. Number of licensed new motor vehicles compared to number of motor vehicles involved in recall for the years 2011 to 2019. Source: Maione et al. (2021b).

4. Machine learning models

Through a study conducted by McKinsey (2020), it can be seen that ML, and AI as a whole, show strong growth within companies in various sectors, with 50% of respondents of the survey stating that they have adopted AI in at least one business function. In addition, 80% of the sample showed growth in profits with the application of Machine Learning techniques and models in the business. Thus, introducing this potential tool into the world of automotive recalls becomes natural and necessary for a greater modernization and updating of the sector.

The study of machine learning is characterized as a section of artificial intelligence, in which data are not given parameters for the resolution of a task by the computer, but access to data. The computer then has the responsibility to, based on this data, find the best way to perform a certain activity to achieve a determined goal alone (Yu & Malan, 2020).

The algorithms can be divided into three main paradigms: Unsupervised learning, Supervised Learning, and Reinforcement Learning. Among these three main categories, the difference comes from the way in which the inputs and outputs are used by the model. Additionally, there are semi-supervised learning algorithms, but these are used on a smaller scale and are a mixture of the unsupervised and supervised learning models, so they will not be applied in this paper, Figure 2.

For the construction and development of the three deprecated models, the scikit-learn library was used. Originally called scikits.learn, it is an open-source machine learning library for the Python programming language and includes several clustering, classification, and regression algorithms including support vector machines, random forests, gradient boosting, k-means, and DBSCAN. To speed up the development of models, it is designed to interact with the numerical and scientific Python libraries, NumPy and SciPy. The project was started in 2007 as a Google Summer of Code project by David Cournapeau. After almost 15 years, the library is one of the most important in the study of machine learning around the world (Scikit-Learn, 2021a).

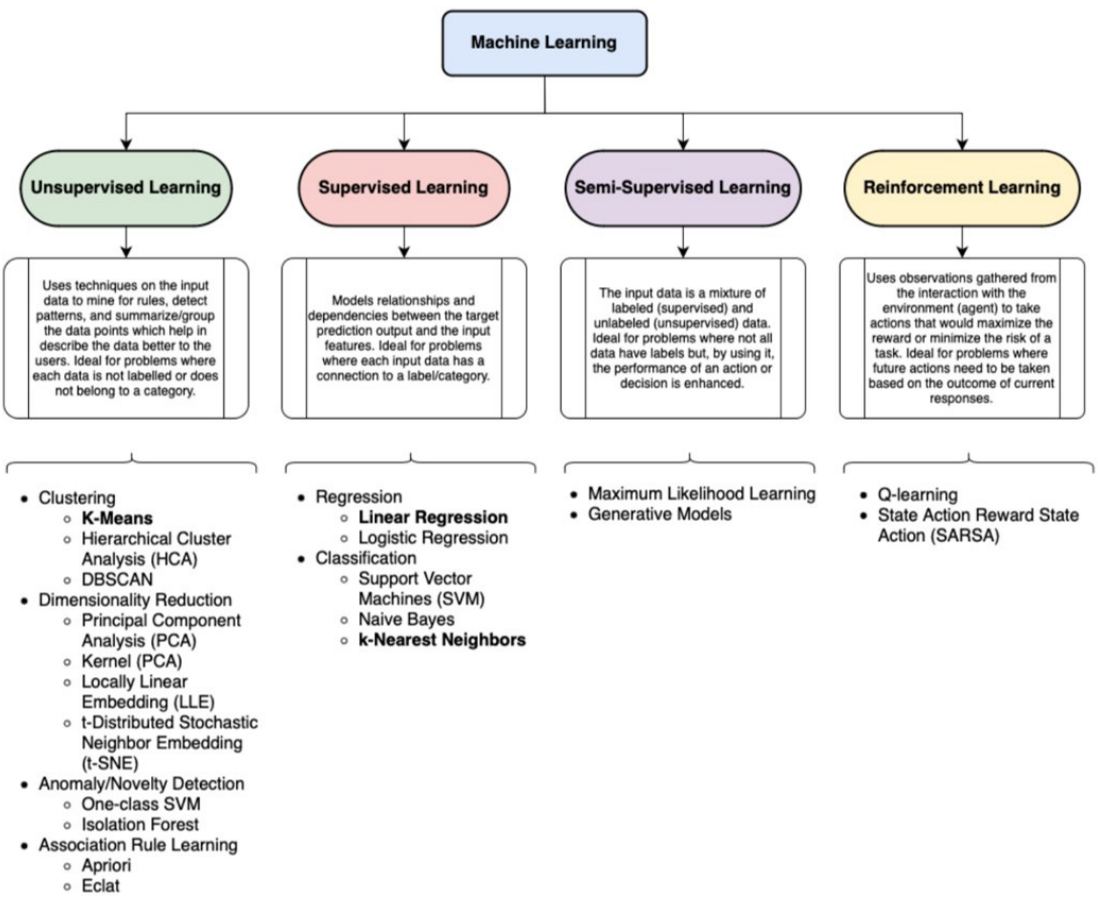

Figure 2. Flowchart of machine learning techniques. **Source:** Adapted from Rafique & Velasco (2018) and Géron (2019).

As a last consideration before diving into the modeling description, it is relevant to mention that all of the following performance coefficients and evaluation processes were performed in testing sets, separately from the training sets used (in all cases where this division was necessary, especially the supervised methods, 70% of the whole data was selected for training and 30% for testing). For more detailed information on the modeling process, see the supplementary material "Modeling Details and Code Availability" (Maione et al., 2023).

4.1. Clustering

Therefore, it begins with the construction of clustering models of the data obtained, that is, of unsupervised learning algorithms, for the problem of automotive recalls. There are several ways to build a model in this way, the one is chosen for this article was the k-means clustering algorithm, as it combines simplicity, allowing a layperson to understand the subject, with efficient results, making it possible to obtain interesting consequences for the sector. In addition, k-means was also chosen since it is typically faster than other common clustering methods, as it does not need to compute distances between data pints for every new data point that comes in, being more scalable and accurate, mixing a simple implementation with convergence guarantee.

Thus, building the model and segmenting the data into three clusters (as justified below), we have the following results shown in Figure 3 (Scikit-Learn, 2021b).

In the European Union, it is not possible to build a model such as the one for the Brazilian database, once the region does not make available the number of affected vehicles call per recall. Thus, it remains a possible study and application to a future study, if EU data are made available, a clustering situation comparing the number of affected vehicles over time, number of recalls over time over time, and a mix of these two variables.

With available data, however, it is possible to build a similar model, compared to the Brazilian case, with USA data, segmenting it into three clusters. Therefore, we have the results shown in Figure 4.

The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1]. Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster (Scikit-Learn, 2021e).

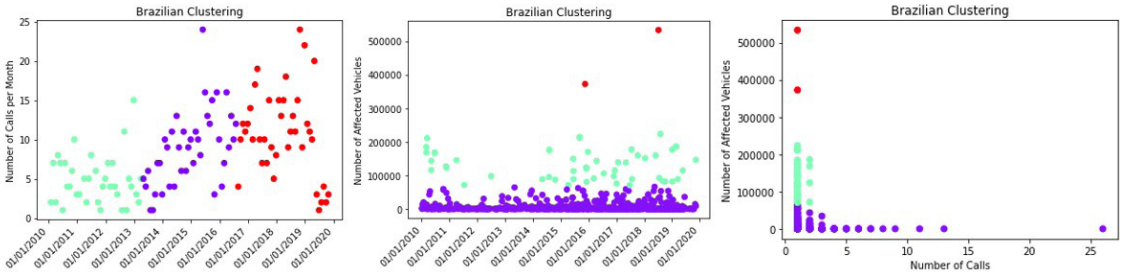

Figure 3. Clustered data of Brazilian Recalls.

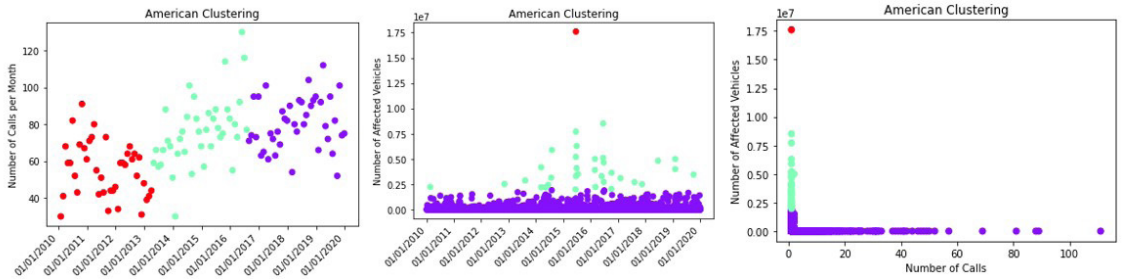

Figure 4. Clustered data of USA Recalls.

Therefore, the intuitive way of choosing the number of clusters is to simulate clustering models with *k* clusters iteratively, with a plausible range for the value of *k,* and choose the value that gives the highest Silhouette Score. As expected by the reader, applying this method to the two studied cases returned the best value of *k* being 2 or 3 and, because of that, in addition to the fact that 3 clusters brought more relevant interpretations to a scientific paper, the graphs and definitive models were built with 3 groups of segmentation, that is, 3 clusters. The silhouette analysis conducted by the authors and the plot of scores, as the whole modeling code, are available in the Git Repository of the project (Maione et al., 2021a).

4.2. Classification

The principle behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point and predict the label from these. The number of samples can be a user-defined constant (k-nearest neighbor learning) or vary based on the local density of points (radius-based neighbor learning), (Scikit-Learn, 2021c).

Neighbors-based classification is a type of instance-based learning or non-generalizing learning: it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point. The k-neighbors classification is the most used technique.

After the process of understanding the model, it is possible to build one for the Brazilian case (Maione et al., 2021a), using two variables: manufacturer and year of manufacture, being the label of interest and the type of risk. The most efficient manner of visualizing this type of algorithm is plotting the variables, if possible (that is if the data has less than four dimensions), and indicating the label of each point, as shown in Figure 5.

It is important to notice that, even though the model has been built with recalls that have started between 2010 and 2019, vehicles from different years before have been part of these recalls, such as the Year Manufacture axis shows.

A similar model can be built for the European Union case, that is, a KNN classificatory using two variables: Manufacturer and Year of Manufacture. However, unlike the Brazilian case, in the European Union, the recalls involved more brands than just the 20 that were present in Brazil, during the time-lapse of interest (2010-2019). To normalize and allow a comparison between both regions, the European Union model was used and selected only the 20 most frequent brands, in terms of the number of recalls, during the period studied. Note that these 20 manufacturers are not the same as the ones used before, on the Brazilian model.

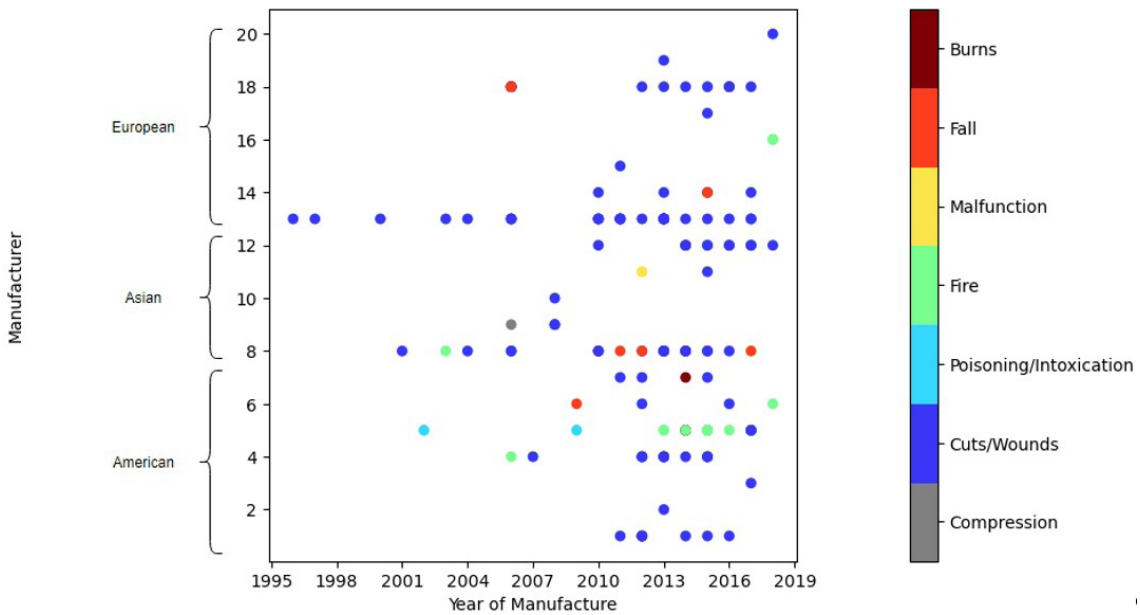

Figure 5. Data distribution of the elements used to build the model of classification.

Another important consideration, involving the European Union case, is that the classification, that is, the labels given in the database followed the patterns of the source. For this reason, the graphic presents hybrid risk classifications, such as Fire/Injuries since the source classified it in this way. Such as the previous location, the most efficient manner of visualizing KNN algorithms is plotting the variables if possible, and indicating the label of each point, as shown in Figure 6.

It is important to notice that in the Brazilian case, even though the model has been built with recalls that have started between 2010 and 2019, vehicles from different years before have been part of these recalls, such as the Year Manufacture axis shows.

Unlike the previous cases, unfortunately, it is not possible to build the same model for the case of the United States since the base provided by ODI does not contain the explicit risk type variable.

The approximate accuracy of each model is given in Table 3, basically showing how often is the classifier correct, once the accuracy of classification models is measured from the proportion of correct predictions made on the test set. This methodology was used in this paper as well.

For registration and future replications by the reader, in the models, after performing cross-validation tests until k = 15, the criteria of 3 neighbors (k = 3) were used, that is, the check for classification is done by observing the 3 closest points, and the precision was obtained by splitting the dataset such as 70% for training (model learning) and the other 30% for the accuracy test itself, as is common in classification algorithms. How well the accuracy (and the classifier itself) represents reality is an important fact to be discussed, given the conditions of the base, and will be done in the results analysis session.

4.3. Regression

Machine learning, more specifically the field of predictive modeling, is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explicability.

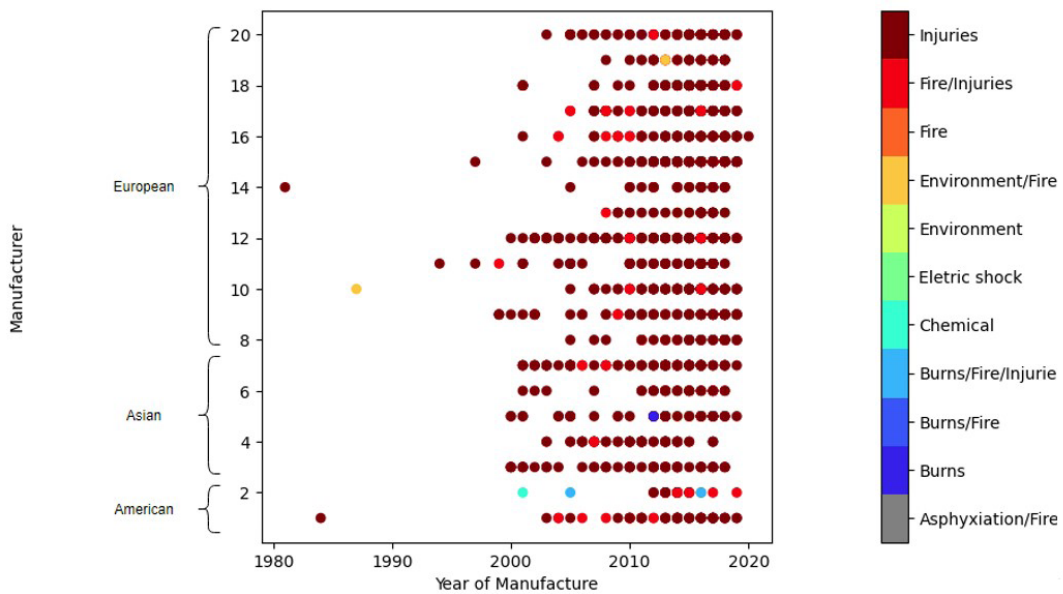

Figure 6. Data distribution of the elements used to build the model of classification.

Table 3. Approximate accuracy of the KNN Algorithm for each location.

KNN Algorithm	Accuracy
Brazil	70%
European Union (EU)	82%
United States of America (USA)	-

Source: Own elaboration.

In applied machine learning, it is common the action of borrowing and reuse algorithms from many different fields, including statistics, and using them towards these ends, achieving scalability, such an important and differential characteristic of artificial intelligence methods. With that in mind, even though regression is a simple statistical method, it is important to apply it considering machine learning approaches and computational speed, to obtain reasonable insights from the used dataset.

Linear regression, the most common and simple method, chosen to be used in this paper, fits a linear model with coefficients $w=\left(w_1,\ldots,w_p\right)$ to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation (Scikit-Learn, 2021d). Mathematically it solves a problem, called Ordinary Least Squares, of the form shown in Equation 1.

$$\min_{w}\left\|Xw-y\right\|_2^2 \tag{1}$$

Equation 01: Ordinary Least Squares

The coefficient estimates for Ordinary Least Squares rely on the independence of the features. When features are correlated and the columns of the design matrix have an approximately linear dependence, the design matrix X becomes close to a singular and as a result, the least-squares estimate becomes highly sensitive to random errors in the observed target, producing a large variance.

Thus, giving the theory of the method, for each region, it is possible to create two basic regressions: one involving the number of recalls per year and another involving number of affected vehicles per year (it is important to remember that the second analysis will only be possible to the Brazilian and the USA cases, because of the restricted European Union database).

Following the algorithm and, one more time, applying Scikit-Learn library, the Brazilian model for the two regressions has been built (Maione et al., 2021a), with results given in Figure 7.

Just like Statistics, in machine learning, it is important to calculate and notice three basic coefficients of regression: the regression coefficient (signifies the amount by which change in x must be multiplied to give the corresponding average change in y, or the amount y changes for a unit increase in x), the mean squared error (measures the average of the squares of the errors, that is, the average squared difference between the estimated values and what is estimated, also denoted as MSE) and the coefficient of determination (gives the percentage variation in y explained by x-variables - the coefficient of determination, R^2, is similar to the correlation coefficient, R, which tells how strong of a linear relationship there is between two variables). All coefficients' methods of calculus (equations) are given by Equations 2 to 4.

$$\beta=\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2} \tag{2}$$

Equation 02: Regression coefficient

$$MSE=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\widehat{y_i}\right)^2 \tag{3}$$

Equation 03: Mean Squared Error

$$R^2=1-\frac{\sum_{i=1}^{n}\left(y_i-\widehat{y_i}\right)^2}{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2} \tag{4}$$

Equation 04: Coefficient of determination

Where $\overline{x}$ and $\overline{y}$ stands for the arithmetic mean of the set of numbers x and y, respectively, and $\widehat{y_i}$ stands for the estimated or predicted values in a predictive model, considering the i^{th} element.

Therefore, Table 4, to organize and show all coefficients of both cases, has been built.

Following the same idea, the EU model for the regression of a number of recalls through the Years has been built (Maione et al., 2021a), with results given in Figure 7. It is important to notice that the construction of the second regression would not be possible, since the European Union does not release the number of affected vehicles on public sources. However, after a request of the authors to the Safety Gate Team of the European Commission, it was possible to get the number of automotive recall's affected items per year, since 2015. To show the greatest information possible, the graph is plotted and shown in Figure 8, but it is important to emphasize that the obtained numbers for this study cannot be compared to the other locations since they represent different time-lapses (Commission's Safety Gate Team, 2021).

Table 4. Coefficients and parameters of the regression with Brazilian data.

Regression - Brazil	Number of recalls	Number of affected vehicles
Regression Coefficient (β)	13.15	2.89E+5
Mean Squared Error (MSE)	988	5.76E+12
Coefficient of Determination (R^2)	0.59	0.11

Source: Own elaboration.

Table 5. Coefficients and parameters of the regression with European Union data.

Regression - EU	Number of recalls	Number of affected vehicles*
Regression Coefficient (β)	42.11	1.48E+06
Mean Squared Error (MSE)	2831	5.43E+13
Coefficient of Determination (R^2)	0.84	0.07

Source: Own elaboration.
*Parameter calculated in a different time-lapse (2015-2019).

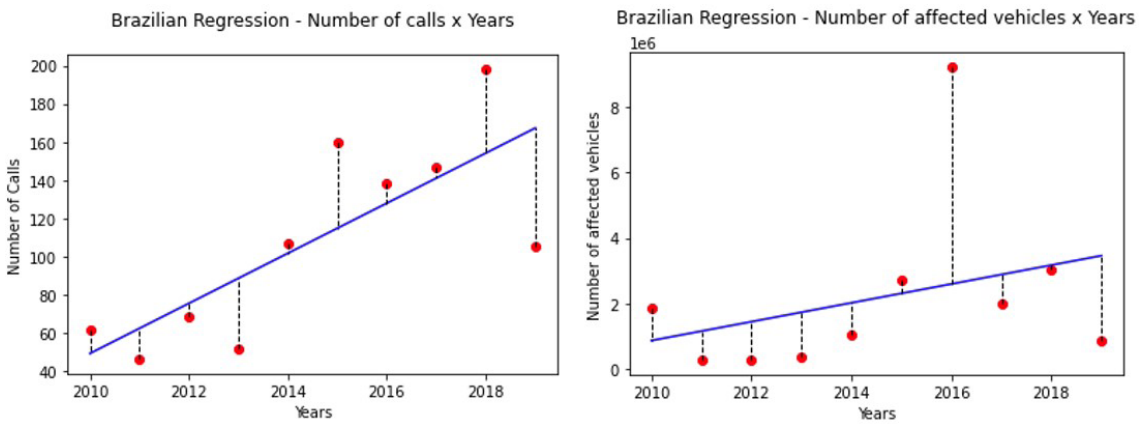

Figure 7. Regression results for both analyses with Brazilian data.

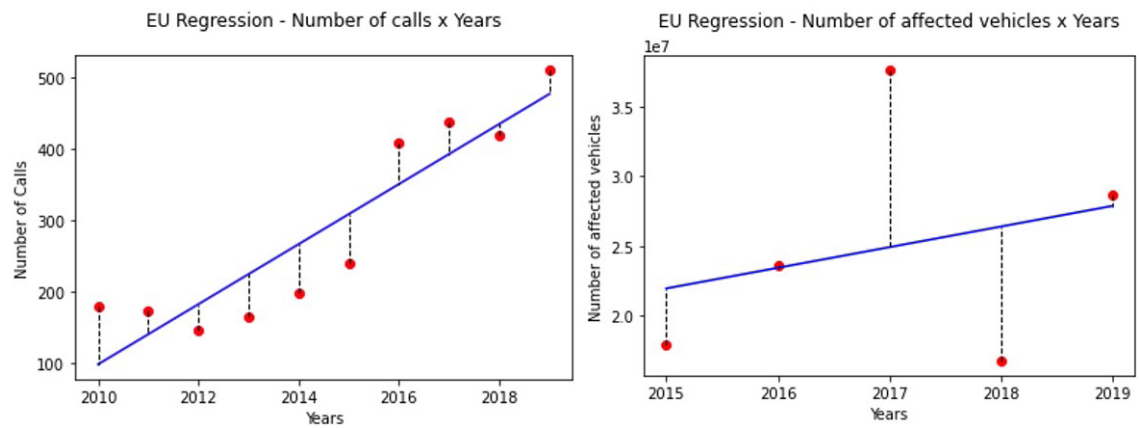

Figure 8. Regression result for the first analysis with European Union data.

In order to organize and show all coefficients of the first case, Table 5 has been built.

Following the same idea, to last, the USA model for the two regressions has been built (Maione et al., 2021a), with results given in Figure 9.

In order to organize and show all coefficients of both cases, Table 6 has been built.

Table 6. Coefficients and parameters of the regression with USA data.

Regression - USA	Number of recalls	Number of affected vehicles
Regression Coefficient (β)	42.79	4.30E+06
Mean Squared Error (MSE)	5448	3.96E+14
Coefficient of Determination (R^2)	0.73	0.28

Source: Own elaboration.

Table 7. Coefficients and parameters of the three regions.

Regression	Brazil		European Union		United States of America	
	Number of recalls	Number of affected vehicles	Number of recalls	Number of affected vehicles*	Number of recalls	Number of affected vehicles
Regression Coefficient (β)	13.15	2.89E+5	42.11	1.48E+06	42.79	4.30E+06
Mean Squared Error (MSE)	988	5.76E+12	2831	5.43E+13	5448	3.96E+14
Coefficient of Determination (R^2)	0.59	0.11	0.84	0.07	0.73	0.28

Source: Own elaboration
*Parameters calculated in a different time-lapse (2015-2019)

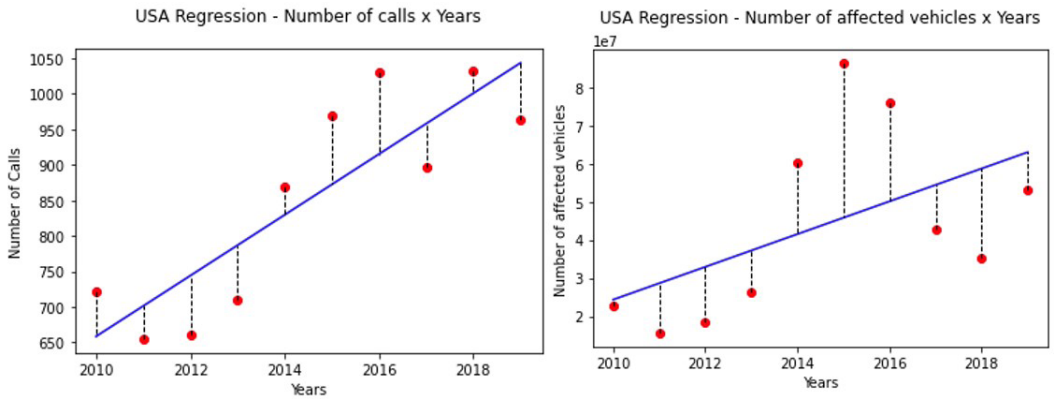

Figure 9. Regression results for both analyses with USA data.

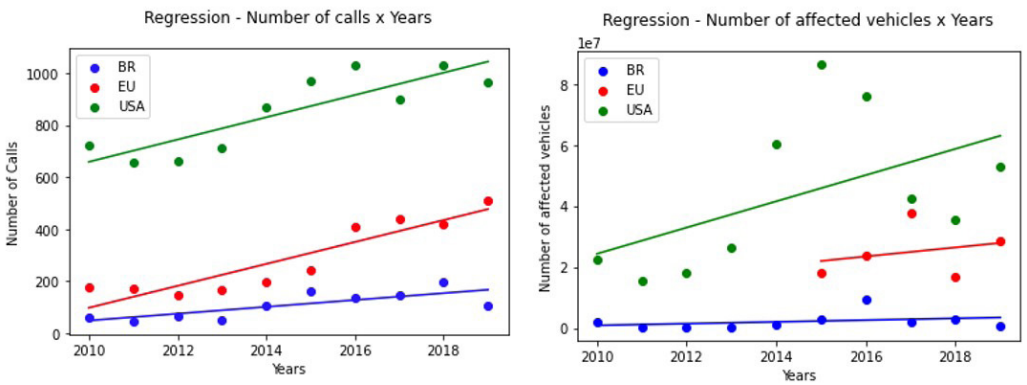

Figure 10. Regression results for the three regions.

As a conclusion from the models' construction, it is possible to build Figure 10 (with all three regressions in one plot) and Table 7 (with all coefficients, from the three locations), without the coefficients from the European Union for the number of affected vehicles, once, as said, it is not fair to use them in comparisons.

5. Results analysis

Considering the obtained results from the Clustering algorithm, both from the Brazilian and USA cases, it is possible to infer the existence of three classes of recalls, one frequent, timeless, and low impact type; a second type, timeless as well but less frequent and with a higher impact (based on the number of affected vehicles) and a third type, rare with an immense impact, affecting an enormous number of products, commonly on the whole world.

In addition, Figures 3 and 4 collaborate on an idea that the regression algorithm confirmed next, as it will be discussed shortly, not only the number of affected vehicles but also the number of recalls, in Brazil and in the USA, are increasing and have grown over the years of the last decade. The first graph obtained for each location analysis testifies to the assumption involving the number of recalls and the second and third graphs obtained for both locations justify the proposal involving the number of affected vehicles, since it is possible to notice a concentration of dots, in the scatter diagram, higher in the last years of the time-lapse of interest.

To last, involving the Clustering results, it is important to highlight the facility that the construction of the scatter diagrams proposed provides an initial understanding and the formulation of an overview of the universe of automotive recalls. That, in addition to a detailed database, made available by the two mentioned locations, allows the creation of new measures of containment and policies that prevent further occurrences.

Other interesting considerations can be made about the second type of constructed algorithm, that is, Classification. Through KNN (k-nearest neighbors) it was possible not only to build a predictive classifier, the main output of the task but also to plot information about recalls made in Brazil and in the EU, involving the manufacturer and the year of manufacture of the vehicle, with risk type labels.

The accuracy obtained in both cases of KNN applications was satisfactory but needs further analysis. It is possible to notice that the Brazilian and the European Union cases have a major risk type label, like one another, that being Cuts/Wounds (Brazilian model) and Injuries (EU model). With that information, which is already valid, the doubt is raised, involving the phenomenon called overfitting. Overfitting is a term used in statistics and machine learning to describe when a statistical model fits very well with the previously observed data set, but it is ineffective to predict new results (Silver, 2013). It is common for the sample to show deviations caused by measurement errors or random factors. Over-adjustment occurs when the model adjusts to these. An over-adjusted model has high accuracy when tested with its dataset, however, such a model is not a good representation of reality and should therefore be avoided.

Since the model built was simple and there are no possibilities of structural errors, this phenomenon generates an interesting question about the occurrence of recalls itself. By these values, it is possible to infer that the recalls that took place during the time-lapse of interest were the majority of one single type, mostly involving direct physical damage (cuts, wounds, and injuries). Once this nuance was perceived by the machine, it is natural that the precision is high since the algorithm will easily get it right if it uses the majority legend forecast (Cuts/Wounds in Brazil and Injuries in the UE).

Other curious points about the two classification models built come from the fact that the European Union involves older cars in its most recent recalls, since the 80s, and, in addition, having most EU brands being the main ones in terms of number of the recalls (that is why they are among the 20 selected). Both the age of retroactive calling cars and the proportion of the locations of the brands are not shown in the same way for the Brazilian case. Among the possible causes are the different laws of the two locations and the intuitive idea that, in Europe, more EU cars are used, therefore the most frequent calling companies will be from the locality itself.

Ultimately, it is possible to make observations about the most intuitive model built, the Regression. As the first impression over the database, the regression coefficient (β) of all cases shows an upward trend, since it is positive, and the obtained regression line is increasing. That fact alone is not surprising and, in a way, is even expected, once the legislation is getting more solid and the internal control of the companies is getting tougher. What is interesting to notice, however, as shown by Baraldi & Kaminski (2019), is that the number of vehicles produced is stable, especially in the European Union and in the USA, due to low annual growth.

Such controversy can be explained by the greater globalization of parts of the models produced in different locations, as occurred in the Takata's airbag and the General Motors' ignition key cases. This fact, if added

to the occurrence of product recalls, where the failure comes from the development and validation of each vehicle, in contrast to the project recall, where the problem is predecessor, in the design phase, contributes to greater growth in the number of recalls and affected vehicles, compared to vehicles produced in the locations of interest.

6. Conclusions

Among the various conclusions that can be drawn from the analysis of the results, the main one can be cited as being the potential that web scraping, machine learning, and technology in general, have to contribute to the understanding of problems that, in principle, do not have plausible and intuitive relationships with these methods. With the web scraping technique, it was possible to obtain an official Brazilian database, more complete and more detailed than any other ready and made available by the responsible body.

The application of the technique and the construction of scripts responsible for scraping, in addition to an extensive search for the bases of the other two regions, the European Union and the USA, followed by the construction of the machine learning models, made it possible to obtain graphics never before seen and with a level of detail and work that make countless reflections possible, some of them already done in this paper, others still to be done in future works. One of the main contributions of this paper, therefore, is the pioneering application of simple Machine Learning models to a field of studies marked by qualitative or simple statistical analyses. It is also possible to mention the fact that ML is inherent to Computer Science and Optimization, which allows faster learning over a large database, which was not a concern of previous statistical analyses and papers. The goal of the technique, and the intention of the application by this paper, is not to come up with a superb new knowledge about the data, but rather with a workable and reproducible model for which the error tolerance is determined by future projects and qualitative interpretations that will be undertaken.

Another important conclusion that can be taken from the whole work of data search is the impact, the importance, and the motives that lead certain locations on making available a group of data and not making a different group, such as the organized database in Brazil, the number of affected vehicles in European Union and the risk type in USA databases. The correct and precise answer of the motive is not possible to be given but can be raised some possible explanations, ranging from the simple option of not grouping that certain data (or inserting it indirectly into another column) to cultural non-disclosure since it is not common, or it is not requested by the public.

For this paper, 12,466 recalls were analyzed among the three locations, and more than 10 versions of the models of interest were built, to get the right and reliable numbers and coefficients, the databases of the regions went through 3 versions each of extra manual treatments for a more intuitive consumption by the authors in the out-of-code environment.

Finally, it is worth mentioning the duality of the explanations in this paper, which tried to be elucidating for the novice reader who does not know machine learning methods, while trying to be challenging to an expert in the subject that seeks to learn about the application of the traditional methods in different areas, such as employees of the manufacturers themselves and the automotive market in general who seeks inspiration for a better interpretation and understanding of the impact of the automotive recalls.

The last conclusion to be drawn from the work is the relevance and magnitude that automotive recalls are taking on in civil society. This fact can be obtained in several ways. In this paper, for example, we have got involuntarily from the three paths, clearly visible on the main three plotted to scatter diagrams. The difference, however, both for the academic community and for the manufacturers, lies in the applied techniques and differentiated perception methods. The perception and recognition of the problem of the increase in the number of recalls, the number of vehicles, affected, and the frequency of high-impact recalls become valuable insofar as they naturally stand out since the applied technique and the arduous construction make the understanding intuitive. An alert of attention, about the magnitude and the real impact of automotive recalls, can be done to the extent that a machine, through mathematical statistical techniques, shows and makes the problem clear, without bias, misunderstanding, or other pitfalls.

References

Alppaydin, E. (2010). *Introduction to machine learning* (2nd ed.). Cambridge: Massachusetts Institute of Technology.

Baraldi, E. C., & Kaminski, P. C. (2016). *A study on the causes of recall in automotive vehicles marketed in Brazil* (Vol. *36*, pp. 1-6). USA: SAE International. http://dx.doi.org/10.4271/2016-36-0169.

Baraldi, E. C., & Kaminski, P. C. (2018). Reference model for the implementation of new assembly processes in the automotive sector. *Cogent Engineering*, *5*(1), 1482984. http://dx.doi.org/10.1080/23311916.2018.1482984.

Baraldi, E. C., & Kaminski, P. C. (2019). The recall and how the lessons learned can be used. In *Proceedings of the 27th International Colloquium of Gerpisa*, Paris.

Baranauskas, J. A. (2011). *Clustering (Agrupamento)*. Departamento de Física e Matemática (FFCLRP-USP). Retrieved in 2020, May 10, from dcm.ffclrp.usp.br/augusto/teaching/ami/AM-I-Clustering.pdf

Bates, H., Holweg, M., Lewis, M., & Oliver, N. (2007). Motor vehicle recalls: Trends, patterns and emerging issues. *Omega*, *35*(2), 202-210. http://dx.doi.org/10.1016/j.omega.2005.05.006.

Brasil. Ministério da Justiça. (2017). *Campanhas de recalls 2017*. Brasília: Secretaria Nacional do Consumidor.

Brasil. Ministério da Justiça. (2018a). Retrieved in 2018, July 12, from http://www.justica.gov.br

Brasil. Ministério da Justiça. (2018b). *Recalls bateram novo recorde em 2017*. Retrieved in 2018, March 19, from http://portal.mj.gov.br

Choi, S. U., Lee, K. C., & Na, H. J. (2022). Exploring the deep neural network model's potential to estimate abnormal audit fees. *Management Decision*, *60*(12), 3304-3323. http://dx.doi.org/10.1108/MD-07-2021-0954.

Commission's Safety Gate Team. (2021). *Motor vehicle cases* [personal message]. Retrieved in 2021, February 21, received by emiliobaraldi@usp.br

Committee on Commerce, Science, and Transportation. (2019). *Danger behind the wheel: the Takata airbag crisis and how to fix our broken auto recall process*. United States Senate. Retrieved in 2019, March 24, from https://www.govinfo.gov

Conner, S. L., & Wanasika, I. (2018). General motors: the ignition switch from hell. *Journal of Case Studies*, *36*(2), 66-81.

Consumer Product Safety Commission - CPSC. (2018). *Consumer product safety commission*. Retrieved in 2018, July 10, from https://www.cpsc.gov

Eilert, M., Jayachandran, S., Kalaignanam, K., & Swartz, T. A. (2017). Does it pay to recall your product early? An empirical investigation in the automobile industry. *Journal of Marketing*, *81*(3), 111-129. http://dx.doi.org/10.1509/jm.15.0074.

European Commission (2020). *Safety Gate: Rapid Alert System for dangerous non-food products - Search alerts*. Retrieved in 2020, August 22, from https://ec.europa.eu/consumers/consumerssafety/safetyproducts/rapex/alerts/?event=main.searchlng=ensearchResults

European Commission. (2018a). *General product safety directive*. Retrieved in 2018, March 19, from https://ec.europa.eu

European Commission. (2018b). *RAPEX - Rapid alert system for dangerous non-food products*. Retrieved in 2018, September 13, from https://ec.europa.eu

European Union. (2002). Directive 2001/95/EC of the European parliament and of the Council. *Official Journal of the European Communities*, pp. 14.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras & Tensor Flow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). Canada: O'Reilly Media.

Gruber, G. E., Kaminski, P. C., & Baraldi, E. C. (2021). A comparison of motor vehicle recalls between Brazil and Germany: different approaches and results. *Product: Management & Development*, *19*(1), e20200003. http://dx.doi.org/10.4322/pmd.2020.035.

Haefele, S., & Westkamper, E. (2014). Identification of product safety-relevant tasks for global automotive manufacturers. *Procedia CIRP*, *17*, 326-331. http://dx.doi.org/10.1016/j.procir.2014.02.052.

Hora, M., Bapuji, H., Roth, A.V. (2011). Safety hazard and time to recall: the role of recall strategy, product defect type, and supply chain player in the US toy industry. *Journal of Operations Management*, *29*(7-8), 766-777. http://dx.doi.org/10.1016/j.jom.2011.06.006.

Imielinski, T., & Mannila, H. (1996). A Database perspective on knowledge discovery. *Communications of the ACM*, *39*(11), 58-64. http://dx.doi.org/10.1145/240455.240472.

Janssen, C., Sen, S., & Bhattacharya, C. B. (2015). Corporate crises in the age of corporate social responsibility. *Business Horizons*, *58*(2), 183-192. http://dx.doi.org/10.1016/j.bushor.2014.11.002.

Kalaignanam, K., Kushwaha, T., & Eilert, M. (2013). The impact of product recalls on future product reliability and future accidents: Evidence from the automobile industry. *Journal of Marketing*, *77*(2), 41-57. http://dx.doi.org/10.1509/jm.11.0356.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine Learning: A review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159-190. http://dx.doi.org/10.1007/s10462-007-9052-3.

Kumar, S., & Schmitz, S. (2011). Managing recalls in a consumer product supply chain – root cause analysis and measures to mitigate risks. *International Journal of Production Research*, *49*(1), 235-253. http://dx.doi.org/10.1080/00207543.2010.508952.

Mackelprang, A., Habermann, M., & Swink, M. (2015). How firm innovativeness and unexpected product reliability failures affect profitability. *Journal of Operations Management*, *38*(1), 71-86. http://dx.doi.org/10.1016/j.jom.2015.06.001.

Maione, B. F., Kaminski, P. C., & Baraldi, E. C. (2021a). *Automotive Recall Codes*. GitHub repository. Retrieved in 2021, December 21, from https://github.com/Maionepy/Automotive-Recall-Codes

Maione, B. F., Kaminski, P. C., & Baraldi, E. C. (2021b). The different legislation of automotive recall and their implications for society. *SAE International*, *36*, 1-12. http://dx.doi.org/10.4271/2020-36-0084.

Maione, B. F., Kaminski, P. C., & Baraldi, E. C. (2023). The automotive recall data search and its analysis applying machine learning [Supplemental material - Modeling Details and Code Availability]. *Production*, *33*, e20220117. http://dx.doi.org/10.1590/0103-6513.20220117.

Maiorescu, R. D. (2016). Crisis management at General Motors and Toyota: an analysis of gender-specific communication and media coverage. *Public Relations Review*, *42*(4), 556-563. http://dx.doi.org/10.1016/j.pubrev.2016.03.011.

McKinsey. (2020). *The state of AI in 2020*. Retrieved in 2021, December 21, from https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020

Medeiros, M. M., & Maçada, A. C. G. (2022). Competitive advantage of data-driven analytical capabilities: the role of big data visualization and of organizational agility. *Management Decision*, *60*(4), 953-975. http://dx.doi.org/10.1108/MD-12-2020-1681.

National Highway Traffic Safety Administration - NHTSA. (2018a). *Toyota Motor Corp. Will pay record $17.35 million in civil penalties for alleged violations of federal law*. Retrieved in 2018, February. 21, from https://www.nhtsa.gov

National Highway Traffic Safety Administration - NHTSA. (2018b). *Recall*. Retrieved in 2018, July. 17, from https://www-odi.nhtsa.dot.gov

ODI. (2021). *NHTSA's Office of Defects Investigation*. Retrieved in 2021, February. 21, from https://www-odi.nhtsa.dot.gov/downloads/

Rafique, D., & Velasco, L. (2018). Machine learning for network automation: overview, architecture, and applications [Invited Tutorial]. *Journal of Optical Communications and Networking*, *10*(10), D126-D143. http://dx.doi.org/10.1364/JOCN.10.00D126.

Recalls. (2018). *Your online resource for recalls*. Retrieved in 2018, July. 10, from www.recalls.gov

Rupp, N. G., & Taylor, C. R. (2002). Who initiates recalls and who cares? Evidence from the automobile industry. *The Journal of Industrial Economics*, *50*(2), 123-149. http://dx.doi.org/10.1111/1467-6451.00171.

Salazar-Reyna, R., Gonzalez-Aleu, F., Granda-Gutierrez, E. M. A., Diaz-Ramirez, J., Garza-Reyes, J. A., & Kumar, A. (2022). A systematic literature review of data science, data analytics and machine learning applied to healthcare engineering systems. *Management Decision*, *60*(2), 300-319. http://dx.doi.org/10.1108/MD-01-2020-0035.

Scikit-Learn. (2021a). *About us*. Retrieved in 2021, Feb. 13, from https://scikit-learn.org/stable/about.html

Scikit-Learn. (2021b). *Clustering*. Retrieved in 2021, Feb. 13, from https://scikit-learn.org/stable/modules/clustering.html

Scikit-Learn. (2021c). *Nearest neighbors*. Retrieved in 2021, Feb. 13, from https://scikit-learn.org/stable/modules/neighbors.html

Scikit-Learn. (2021d). *Linear models*. Retrieved in 2021, Feb. 14, from https://scikit-learn.org/stable/modules/linear_model.html

Scikit-Learn. (2021e). *Selecting the number of clusters with silhouette analysis on KMeans clustering*. Retrieved in 2021, March 3, from https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Secretaria Nacional do Consumidor - SENACON. (2020a). Retrieved in 2020, November 27, from http://portal.mj.gov.br/recall/principal/openData

Secretaria Nacional do Consumidor - SENACON. (2020b). *Alertas de Recall*. Retrieved in 2020, November 27, from http://portal.mj.gov.br/recall/

Silva, P.B., Andrade, M., Ferreira, S. (2020). Machine learning applied to road safety modeling: a systematic literature review. *Journal of Traffic and Transportation Engineering*, *7*(6), 775-790. http://dx.doi.org/10.1016/j.jtte.2020.07.004.

Silver, N. (2013). *O sinal e o ruído*. Rio de Janeiro: Editora Intrínseca.

Slack, N., Chambers, S., & Johnston, R. (2010). *Operations management* (6th ed.). London: Pearson Education.

Wakefield, K. (2020). *A guide to machine learning algorithms and their applications*. UK: SAS.

Yu, B., & Malan, D. J. (2020). *CS50 introduction to artificial intelligence with Python*. Retrieved in 2020, December 15, from https://cs50.harvard.edu/ai/2020

Zhu, A. Y., von Zedtwitz, M., & Assimakopoulos, D. G. (2018). Responsible product innovation: putting safety first. In E. G. Carayannis (Ed.), *Innovation, technology, and knowledge management*. Switzerland: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-68451-2.

Supplementary Material

Supplementary material accompanies this paper.

Modeling Details and Code Availability

This material is available as part of the online article from: https://doi.org/10.1590/0103-6513.20220117