Vol. 10(1), pp. 9-18, January-June, 2024
DOI: 10.5897/IJCER2023.0310
Article Number: 79D45BE71985
ISSN: 2141-6494
Copyright ©2024
Author(s) retain the copyright of this article
http://www.academicjournals.org/IJCER

International Journal of Computer Engineering Research

Full Length Research Paper

Lessons learned and perspectives on constrained data collection and preparation for a predictive machine learning model applied to transportation industry in a non-digitalised environment

Simon Isaac KABEYA MWEPU[1]* and Patrick MUKALA[2,3]

[1]Higher Institute of Statistics of Lubumbashi, Democratic Republic of Congo.
[2]University of Wollongong in Dubai, United Arab Emirates.
[3]National Pedagogical University, Democratic Republic of Congo.

Received 28 October, 2023; Accepted 15 January, 2024

Machine learning algorithms are based on qualitative and quantitative historical data, to create predictive models for shape recognition, autonomous systems, etc., using classifiers like K-Nearest Neighbors (KNN), neural-network, etc. So, treatment of data is the undisputed fuel that powers any machine learning endeavour. A standard data collection methodology would comprise a few steps as data collection, cleaning, resampling, resizing, selecting variables, extracting features, transforming and projecting data, removing noise, and irrelevant information. In this paper, we report on a case study based on collection of data for predicting trains derailments in the context of spiking neural networks (SNNs), the rail carrier in the Democratic Republic of Congo. We share the lessons learned of a company, where pretty much everything is done manually on reliance on experts' opinion. Our data collection approach at SNCC, concerns 117,473 vehicles including 15,280 derailed of which 111 come from networks outside SNCC. 25,727 vehicles were excluded for one of the reasons mentioned earlier. The remaining 86,463 vehicles were split into 2 blocks consisting, respectively of 69,170 vehicles for the learning data and 17,293 vehicles for the test data. KNN classifier predicts the (survenue) of derailments with 87% for 3-NN and 85% for 3-NN of rate. With this rate, it is possible to avoid derailments by predicting their (survenue). But we must perform it to avoid consequences of derailments on persons and materials.

Key words: Machine learning, K-nearest neighbors (KNN), neural-network, spiking neural networks (SNNs), data constraints, predictive maintenance, train vehicles.

INTRODUCTION

Machine learning algorithms build predictive models based on historical data. These algorithms are used to

*Corresponding author. E-mail: kabeyamwepu@gmail.com. Tel: +243970058613.

figure predict new outcomes that humans did not have access to before the operation was performed. These results are usually new data (Lantz, 2019; Agrawal, 2019; Abaker and Saeed, 2021; Kamalov et al., 2021; Arumugam et al., 2022; Nordin et al., 2021). Knowing that in Machine Learning technics, the quality and the quantity of data are very important factors for the construction of any model. No learning can provide acceptable results without data that meets the requirements qualitatively and quantitatively (Siebert et al., 2021; Wang et al., 2021; Toye et al., 2014; Cai and Zhu 2015). The two statements show that data plays an important role in setting up a Machine Learning model. Moreover, concrete applications of Machine Learning using data are more and more numerous, lately. Among the most used models are predictive analytics, pattern recognition, autonomous systems, conversational systems otherwise known as hyper-personalization, etc. All these applications have one thing in common: they start from the good understanding of an operational problem attached to the data and algorithms to be used (Szepesvári, 2022; Alzubaidi et al., 2021; Beer, 2019).

The method of setting up Machine Learning applications, starting with the CRISP-DM method was developed 25 years ago and which had only one version, up to the Cognitive Project Management for Artificial Intelligence (CPM-AI) currently used, given a large part to data processing (Adams, 2023; Kolyshkina and Simoff, 2021; Ayele, 2021; Schröer et al., 2021; Ribeiro et al., 2020; Studer et al., 2021). Data plays a very important role in the realization of Machine Learning projects. Their treatment, regardless of the method chosen, constitutes the second most important point after the <understanding of the problem or the profession> (Menkel-Meadow, 2017). Understanding data is necessary because understanding the business or approving the project is not the starting point for building the Machine Learning model. But on the contrary, you need to have an appropriate set of data to deal with the problem. Knowing that a Machine Learning model is typically built by learning and generalizing from a set of training data, and then applying valuable information obtained to new data to make a prediction (training, validation and testing).

The lack of data at the outset is a real reason why a Machine Learning model is not being implemented. In other words, without data, there is no possibility of setting up a Machine Learning model (Vollmer et al., 2020); Arnal et al., 2022; Heinrichs and Eickhoff, 2020; Chekroud et al., 2021; Chen et al., 2023). But, access to data alone is not a guarantee of obtaining a good Machine Learning model, we need cleaned data, of good quality and sufficient in quantity to complete a Machine Learning project. This last statement tells us that: (1) the data processing step is crucial and cannot be skipped and (2) it is necessary to identify your needs in terms of data in order to see if the games are complete and of sufficient quality to carry out a Machine Learning project; so, wanting to set up a predictive analysis model in Machine Learning for the anticipatory announcement of the occurrence of random events such as derailments or traffic accidents. What operations do we need to apply to ensure that the data obtained is an effective set of information that can lead us to a better prediction?

SOLUTIONS

Methodology applied

The following procedure were applied: (1) data collection; (2) data cleansing; (3) selection of predictive variables (by the KHI-SQUARE procedure); (4) evaluation of the importance of each variable (by Multiple Discriminant Analysis); (5) modelling by the KNN method.

Application of the procedure

Achieving a consistent, efficient, responsive and better set of data quality requires an adapted approach to data processing. As shown earlier, the data processing step cannot be skipped because it is crucial and necessary for the identification of your needs in terms of data that can eventually lead to a Machine Learning project.

The realization of these two statements leads us to answer several questions relating to the data to be used.

Understanding and identifying the data to be used

On this point, we retain some of the following essential questions:

Where can we find the training data for our Machine Learning model? Generally, old companies (such as SNCC) that have made time (more than 25 years) have recorded their data on paper files for archaic processing. Such files require long periods of input to the computer, transformation in order to adapt them to the project to be set up and a selection of values to be used, because usually such files contain a lot of data, some of which is not useful for the project to be developed. These are often files containing a lot of information, much of which is unrelated. These data coexist in these databases by the fact of coming from the same company. The great advantage of such files is that they give the possibility to find all the elements in the different files relating to the history of the company. Unfortunately, their great disadvantage is that they need to be entered on the computer. This can easily take you 10 months depending on the size of the information to be entered. As an example, we had to enter maintenance information from 117,363 SNCC wagons for 48 explanatory variables. This operation took us 8 months using 2 part-time operators at a rate of 3 h per working day. The correction operations took us 45 working days, with a finding that the data entered at the beginning of the operation when the operators were not yet tired, had few errors unlike the data entered at the end where many errors due to fatigue, negligence and especially habit. Whenever a piece of data starts with "h" for example, operators automatically filled in with "g" to get the abbreviation of wagon, whereas sometimes there was "v" for passenger car or "l" for locomotive, in SNCC's jargon. Other companies older than 10 years have recorded their data on Microsoft-Excel spreadsheets. In such cases, the companies concerned have recorded the data that were previously on paper in Excel tables (case of several departments of SNCC having been recovered in its time by SIZARAIL). Such files require first the selection of the fields concerned, secondly, the transformation by appropriate modules to make them suitable for processing by modules in Python language. Finally, after these two operations, it is necessary to consider merging the two files in order to obtain a single file for training.

In short, data is generally found in companies for which we want to set up a Machine Learning model. On the other hand, there are sites offering data that can be used for Machine Learning projects. But the data from such sites is often test data, which helps us perfect the algorithm we want to implement.

How much data is needed for this Machine Learning project? Many data are suitable for a good training of the methods to be applied. However, for companies such as SNCC, where seizures have been made for long periods, the quantities of data available are sufficient because they relate to several years. Unfortunately, these data were entered without a specific project, they relate to a set of activities revolving around railway operations. That is how we find in this group a little bit of anything. This sometimes unnecessarily inflates the files giving the impression of having a lot of data while the amount of work required to make them usable is exponential.

What is the current quantity and quality of training datasets? In companies such as SNCC, quantity is generally not a problem, but quality leaves something to be desired. Data has often been negligently captured by workers unaware of its importance and above all, seeing only piles piling up without anyone asking anything about the data entered over the years. In addition, the data entered relate to all the activities carried out in this railway company. However, rail data cover many elements, generally according to their own laws. We know that in this regard activities such as the occurrence of incidents follow a fish law (law of rare phenomena) while others such as passenger traffic follow the exponential law (with occurrences increasing as time progresses). Grouping such activities in a single table is commonplace in railway undertakings such as SNCC. Answering the question about quantity will surely pose problems insofar as on the one hand we will have a lot of data and on the other, we will have very little. This will result in the presence of many empty boxes inside the distributions.

In terms of quality, archaic data has a lot of information captured without a single purpose, so it will typically provide most of what a machine learning study is looking for. Such an assembly of data will generally provide the possibility of carrying out several studies, both useful for different treatments using machine learning techniques.

How will the data be distributed between the test set and the training set? Generally, for Machine Learning projects, the data is divided into three parts: 70% for training data, 20% for data for validation, and 10% for test data. Others, on the contrary, propose to divide into two groups of: 80% for training data and 20% for test data. The second approach seems easy because it increases the number of test and training data. For historical data from rail companies like SNCC, these numbers are easy to reach. But the problem often arises when the data of the groups come into play, because many of the data have gaps due to the cohabitation of several laws in the groups of available data. Data cleansing is then necessary in order to retain only the data that can help us draw appropriate conclusions.

For supervised learning tasks, what means are available (financial, technical or human) to label this data? Data labeling is the process of labeling the data of the variable explained (Sarr et al., 2021). Regarding the historical data of the SNCC, the labelling does not pose any problem because coming from several files, the data of the variable explained are well identified as well as those of the explanatory variables; the problem that arises only when merging the different files.

Can pre-trained models be used? Yes, existing models can be used to pre-train. But the problem is that the available data has a lot of voids that must either be removed or treated differently. By the technique of removing individuals with empty values, the number of data may become minimal and detrimental to learning insofar as it loses some valuable information, useful to the conclusions to be taken. Another technique consists in replacing the gaps either by 0, for numerical values and by "NA", for character values, this technique seems good insofar as it avoids losing useful information, but in the end, it does not help insofar as during classification, we often find ourselves with classes consisting only of individuals with these deficiencies (Merriam and Grenier 2019; Guo et al., 2020). These classes are not useful for processing, because they do not give us any important information.

We applied both techniques and found that with the first we lost ±76,000 cars (individuals) that could provide us with useful information to document decisions on whether or not derailments occurred (Kabeya, 2021). The remaining data (41,363 wagons) seemed to us insufficient, in quantity, for the correct treatment, because the predictor used gave an accuracy of only 61%. By applying an intermediate method of removing individuals with many voids and keeping after replacement by "0" or "NA" only the variables of individuals with less than 3 voids, we found a loss of ±31,000 individuals increasing our base from 117,463 wagons (individuals) to 86,463 individuals with 71,193 vehicles not derailed against 15,270 derailed vehicles. Regarding explanatory variables, 4 variables with more than 30% gaps have been removed from the database, increasing the number of explanatory variables from 48 to 44. This second way of proceeding allowed us to increase the accuracy rate from 61 to 87% for 3-NN and 86% for 5-NN (Kabeya, 2021). These results were obtained with our own program in which we replaced the Euclidean distance by the Manhattan distance, since the latter is simpler and more adapted to the case we were treating. In this way, the pre-trained model gave results of 85% for 3-NN and 83% for 5-NN; while our program gave 87% for 3-NN and 86% for 5-NN (Kabeya, 2021). Regarding the study with the fuzzy data, the photos and videos of the defects on the track will be recorded in PNG files and the measurements related to these defects will be filled in Microsoft-Excel files.

Where is the operational data? The operational data is located in databases in Microsoft-Excel, since initially some of it was on paper files (towed vehicle maintenance data sets and derailment data). The rest of the data, that is, load, customer and other data, was in a database managed by MySQL. From this experience, we believe that incident files will always end up in files that are often entered manually. For most cases, these data will come either from the road traffic police services for road transport, or from the safety and/or track maintenance services for rail transport.

Are there ways to access real-time data, edge data (sensor or other), or important data that would be "isolated"? For studies of interview data, the sensors were useless because the data was entered into handwritten files. But for track defect data, sensors are needed to capture defect images so that these photos or videos are compared to ordinary tracks so as to detect errors that may occur during the passage of trains. The data thus collected must generally work offline in collusion with those from the internet and which will be updated regularly as new arrivals occur. The classifier will work in deferred time since learning will take place each time a new batch of information arrives at the terminal. But it will also be able to work in real time in measurement with high performance requirements to provide instantaneous results. The model will be trained multiple times as new results arrive off trains.

Data collection and preparation

This is the most important step in the process of setting up a Machine Learning model. It includes collection, cleaning (“cleansing”), aggregation, augmentation, labelling, standardization and transformation as well as all activities around structured, unstructured and semi-structured data.

The collection aims to collect data from the sources previously mentioned in the previous step. The collection is done only from the appropriate well-identified sources so as to avoid duplication which would come when collecting the same information from several different sources. It is preferable to identify beforehand the sources from which to collect the data. Refer to it for the identified data. The other sources can be used as a source of control or complement, if the data can have certain easily correctable biases by comparing the same files constituted in different working blocks.

This comparison will replace incorrect data with those that reflect the pure reality observed during data collection. Sometimes, it happens that the data collected fails to reach a sufficient quantity (which is not the case for us). In this case, it is necessary to generate data. This solution is only used in cases of extreme necessity or if it is impossible to collect more data. To use generated data, synthetic data generation techniques must be used, which proceed as follows:

(1) Swap existing data
(2) Use a statistical law "close" to that governing existing data
(3) Use a probabilistic model (Monte Carlo method, for example)
(4) Use a deep learning technique: Variational Autoencoder (VAE) or Generative Adversarial Network (GAN)

These four techniques can only be used in extreme emergencies.

Naturally, when the data becomes numerous and coming from various sources; therefore, they have different types. In this case, we prefer to use a specific structure that takes into account all the data of different types (digital data, analog data, characters, photos, video, etc). The current structure fulfilling all these conditions is the Data Lake, which is a large set of raw data, structured or not, where different users would come to examine, scrutinize the data or extract samples from it in order to carry out analyses or identify trends (Boehm et al., 2022; Beheshti et al., 2022; Rehman et al., 2022; Madera, 2018). In the present case, the data being all numeric, character or Boolean, it was not necessary to set up a data lake. But, when we are going to seek to mount a lake containing the faults of the way which will incorporate photos and videos which are fuzzy type data, in addition to data on the feelings of humans which, also, are fuzzy data, we will be called upon to build a structure in the form of a data lake in order to take advantage of the following advantages:

(1) Support for data in native form: low cost
(2) Ability to store a wide variety of data types and formatting at the time of support: schema on read
(3) Perform analyses based on a single domain, such as the initially fuzzy value
(4) Governance strategies for configuring, identifying, reusing and disposing of data, indicating the origin of the data and the people who handled it and the level of modifications (Wang et al., 2021).

For our case of prediction of derailments, the collection of data from SNCC (without constitution of a Data Lake) provided the variables listed in Table 1.

After data collection, it needs to be cleaned. Cleaning is done by:

(1) Standardization of formats, at this stage, it is necessary to put the data in formats such that they can be easily processed by the data preparation software.
(2) Data resampling (feature scaling) to modify the data distribution, if necessary. This step is important in the case where the data is unbalanced (over-sampling or under-sampling will make it possible to best represent all the data)
(3) Resizing of variables which consists of putting all the features on the same numerical scale. This is what constitutes the normalization or standardization.
(4) Selection of variables (Features selection) which consists in arranging the variables in order of importance, in our case the MDFA, allowed us to carry out this arrangement. After this arrangement, we will obtain new variables if we use the PCA.
(5) Features engineering which will consist of adding new data features that can help you better predict the data. Generally, these new characteristics come from the decomposition into several new values of the old attributes or from the decomposition into main factors of the old variables if PCA has been used.
(6) Transform the data: As many Machine Learning algorithms work best with numerical data having a Gaussian distribution, it is necessary to transform the variables using an exponential function to better expose the data to a learning algorithm
(7) Project the data: Input variables must be made non-generalizable to avoid overfitting. This is done by using a clustering or unsupervised projection method to project he data into a lower dimensional space to reduce model overfitting.

After cleaning, the most important step is to improve the data or improve the sample database. At this stage, several tasks will be applied, among others:

(1) The suppression of noise and superfluous information. In this step, we fit the data into intervals beyond which the data is considered inaccurate. For each explanatory variable, we calculate the arithmetic mean or the median depending on whether it is continuous or discrete data, we calculate the standard deviation and the confidence interval according to the case at 95% or 99%. All the data which will be found outside the interval thus calculated will be considered as erroneous and likely to be replaced, either by the mean or median data, or by a neighboring data. This replacement is acceptable only if the amount of data to be replaced is less than an acceptable rate. If not, it is best to leave the distribution intact.
(2) After noise removal, the data irrelevant for training should be removed for better prediction result. A variable is relevant for prediction when it is related to the explained variable. All predictive variables independent of the predicted variable must be deleted because they do not contribute anything to the improvement of the prediction. Generally, the degree of connection is evaluated by means of the correlation coefficient whose formula is:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \qquad (1)$$

where r = the linear correlation coefficient, S_{xy} = covariance between Xi (the predictor variable) and Y (the predicted variable).

Its formula is:

$$S_{xy} = \sum_{i=1}^{N} X_i Y_i - \frac{\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{N} \qquad (2)$$

where S_{xx} = the variance of the variable X.

Its formula is:

$$S_{xx} = \sum_{i=1}^{N} X_i^2 - \frac{(\sum_{i=1}^{N} X_i)^2}{N} \qquad (3)$$

where Syy = the variance of the variable Y.

Its formula is:

$$S_{yy} = \sum_{i=1}^{N} Y_i^2 - \frac{(\sum_{i=1}^{N} Y_i)^2}{N} \qquad (4)$$

The correlation coefficient evaluated by means of the KHI-SQUARE independence test, whose formula is

$$\sum_{i=1}^{N} \frac{(O_i - t_i)^2}{t_i} \qquad (5)$$

having O_i as observed frequencies of class i and ti as

Table 1. Extract from the coding of variables table.

No.	Variable	Code	Meaning	Modality
1	Network	ResN	The name of the country to which the wagon belongs	1.(86.463), 2.(5.635), 3.(12.579), 4.(3.411), 5.(9.27)
2	No-Derailed/ Derailed	N/D	Non-derailed vehicle or derailed vehicle	N(102.082) and D(15.280)
3	Train composition date	Comp-Date	Day we dialed the train, these dates will be either dry season or rainy season	SS(46.645) and SP(70.717)
4	Train Number	Train_ Num	The train number, these numbers will designate either the Goods, Passenger or Works trains	TM(105.887), TV(9.489) and TX(1.989)
5	Driver name	Driv_ Nam	Name of the train driver, it was taken in the sense of the rank of drivers, namely: Machinist Master Instructor, Machinist Instructor and Line Driver	MMI(2.773), Ins(11.586) and NR(103.003)
6	Number of locomotives	Nbre_ loco	Number of locomotives pulling the train	1(114.613) and 2(1.218), les NR(1.531)
7	Departure station	Sta_dep	The departure station of the train which will be considered taking into account the SNCC regions to which the station belongs	RS(83.796), RC(32.960), NR 633
8	destination station	Sta_des	The destination station of the train which will be considered taking into account the SNCC region to which the station belongs	RS(84.505), RC(32.105), RN(16), GL(0), FR(0) and NR 736
9	Date of departure	Date_ dep	Day when the train left the departure station, will be considered taking into account the seasons	DPSP(72.369) and DPSS(44.855)
10	Departure time	Depart_ time	Evokes the departure time of the train, will be considered taking into account the night hours (7:00 p.m. to 5:59 a.m.) and the daytime hours (6:00 a.m. to 6:59 p.m.)	Day(64.938) and Night(52.289), NR(138)
48	Coupling height 2	Height_ coup2	The height at which the wagon is coupled (back)	<= 879(24.618), >875(879) and NR(58/062)
49	Date of maintenance	Maint_ Date	The date when the maintenance was carried out by season	SSE(24.618) and SPE(34.682) NR(58.062)

Source: Collection from SNCC.

theoretical frequencies of class i.

The application of Equation 1 to the collected data provides us with the correlation coefficients between the predicted variable and the predictor variables. The application of Equation 5 to the table constructed by splitting for each variable, the elements concerning the derailed part and the non-derailed part, we obtain the calculated KHI-SQUARE value. These computed KHI-SQUARE values will be compared to the theoretical KHI-SQUARE value, to obtain the following decisions about the predictor variables (Table 2).

When we applied these formulas to the derailment data, we found that three predictor variables were independent of the predicted variable (the Boolean variable derailment or non-derailment). We excluded the relevant variables from the study. To reinforce the improvement of the quality of the prediction, we prefer to use a general classifier that will help us to retain the variables that gravitate around the predicted variable. We had the choice between PCA (Principal Components Analysis which would provide us with the principal components describing the link between the predicted

Table 2. Table of the results of the KHI-SQUARE test applied to the variables.

No.	Variable	Numbers of modalities	KHI-square calculated	Theoretical KHI-square	Decision
1	Network	5	5.954	9.488	Dependency
2	Comp_Date	2	244.428	3.841	Dependency
3	Train_num	3	672.466	5.991	Dependency
4	Driver_Nam	2	1.489	3.841	Independence
5	Nbre_loco	2	14.311	3.841	Dependency
6	Stat_dep	2	280.684	3.841	Dependency
7	Stat_dest	2	258.78	3.841	Dependency
8	Nbre_véh	5	146.637	9.488	Dependency
9	Dapart_date	2	206.819	3.841	Dependency
10	Depart_time	2	3.813	3.841	Independence
11	Net_weight	5	305.714	9.488	Dependency
12	Hg_ Usage	22	5842.739	31.671	Dependency
13	Hg_ Type	14	4058	22.36	Dependency
14	Trans_Date	2	206.771	3.841	Dependency
15	Com_code	4	732.847	7.815	Dependency
16	Load_weight	4	868.301	7.815	Dependency
17	Num_ric	Is only concerned in the event of a derailment			
18	Incid_Date	Is only concerned in the event of a derailment			
19	Begin_Sect	Is only concerned in the event of a derailment			
20	End_Sect	Is only concerned in the event of a derailment			
21	Cum_S_Km	Is only concerned in the event of a derailment			
22	Cum_S_m	Is only concerned in the event of a derailment			
23	Cum_E_Km	Is only concerned in the event of a derailment			
24	Cum_E_m	Is only concerned in the event of a derailment			
25	Hl_Num	Is only concerned in the event of a derailment			
26	Véh_Loc	Is only concerned in the event of a derailment			
27	Kind_prod	Is only concerned in the event of a derailment			
28	DerDriv_Nam	Is only concerned in the event of a derailment			
29	Previs_caus	Is only concerned in the event of a derailment			
30	Type_Maint	3	54.982	9.991	Dependency
31	Num_ess1	10	555.877	16.919	Dependency
32	Diam_ess_1	3	52.574	5.991	Dependency
33	Num_ess_2	12	518.247	19.675	Dependency
34	Diam_ess_2	3	48.905	5.991	Dependency
35	Num_ess_3	10	504.956	6.919	Dependency
36	Diam_ess_3	5	40.773	9.488	Dependency
37	Num_ess_4	11	519.898	18.307	Dependency
38	Diam_ess_4	4	211.754	7.815	Dependency
39	Num_cyl1	6	177.407	11.070	Dependency
40	Num_cyl2	3	346.49	5.99	Dependency
41	Skate1_game	11	159.84	18.307	Dependency
42	Skate2_game	9	730.043	15.507	Dependency
43	Skate3_game	13	342.89	21.026	Dependency
44	Skate4_game	18	454.38	27.587	Dependency
45	Bogie_Type	15	701.067	23.684	Dependency
46	Hitch_height1	2	5.4	3.845	Dependency
47	Hitch_height2	2	34.737	3.845	Dependency
48	Maint_date	2	2.836	3.845	Independency

Source: Author’s calculations.

variable and the predictor variables). This description would provide the ability to get new variables explanations of the phenomenon under study (in our case the derailment). But our goal is not to identify new explanatory variables, but to observe how the predictive variables cluster around the predicted variable; hence the ACP, seemed to us not useful in this case and we have retained MDFA (Multiple Discriminant Factor Analysis). This classifier will have the merit of giving the image of the difference between the predictive variables and the variable predicted by quantifying this deviation by a generalized distance of MAHALANOBIS (Hasanzadeh et al., 2017; Kasongo, 2003). It also offers the possibility to determine the discriminating power of each predictive variable. The formula is

$$\mu_k = \frac{d_k' G D^{-1} G' d_k}{n d_k' d_k} \tag{7}$$

or

$$\mu_k = \frac{Var(y)}{Var(d_k)} = R^2(yd_k) \tag{8}$$

because $\frac{1}{n} d_k' d_k$ = total variance (Var(d$_k$)) and within-group variance = $\frac{1}{n^2} d_k' G D^{-1} G' d_k$ for Equation 8.

The MDFA applied confirmed the results of the KHI-SQUARE test. So, the selected variables will correctly contribute to the best prediction of the phenomenon under study which is the classification of wagons either in wagon to derail or wagon not to derail. With the MDFA, we have no longer need to identify the important dimensions in order to reduce the treatment dimension to the most important predictor variables; which is what the ACP would have given us (Johnstone et al., 2021; Meng et al., 2020; Kowsher et al., 2019). As the method provides a general description of the predictive variables facing the predicted variable with distance evaluation, we found better to do not reduce the dimension. This method applied to data from the SNCC provided the following information:

(1) First block: technical variables only
(2) Second block: the variables sampled when the derailments occur
(3) Third block: the other variables

Halfway through the three blocks, there are two variables (Hg_Usage (Usage_Hg) and THGL (Hg type)) which do not contribute to the construction of any of the two axes of the foreground, and therefore to the prediction. These groups show that the gradation of the variables to illustrate the role of each variable in the construction of the axes starts from the technical variables towards the other variables, passing through the variables that only come into play when the incident occurs.

This method was applied separately for SNCC wagons and for foreign wagons circulating on the SNCC network was intended to ensure that the maintenance carried out outside the SNCC workshops gives different values from those achieved in the SNCC workshops. Among the 15,280 foreign wagons which circulated on the SNCC network during the period under study, only 111 Mozambican wagons derailed on the SNCC network. By the KHI-SQUARE test, we ensured that the number of derailed SNCC vehicles was different from that of Mozambicans' derailed vehicles; what has been proven. The 111 Mozambican wagons derailed on the SNCC network does not represent much for them to be incorporated into the study. The applied KHI-SQUARE test proved that in the face of 15,280 foreign wagons in circulation on the SNCC network, 111 wagons are only a small negligible minority, which can be excluded from the study without prejudicing the results to be obtained. We excluded them from the study. In this way, we left with only one sample, that is made up of the SNCC wagons.

Having had the sample made up only of SNCC wagons, we will cut it into two blocks, the first made up of 80% for the learning base, which gives a number equivalent to 69,170 wagons and 17,293 wagons for the test base (20%).

The internal composition of these two bases is as follows:

(1) From 86,463 wagons under observation, 15,170 derailed, that represented 17.5%, 71,293 wagons have not derailed, that represented 42.5%.
(2) From 69,170 wagons in the learning base (80% of 86,463), we will retain 57,034 wagons which have not derailed against 12,136 wagons which have derailed.
(3) From 17,293 wagons in the test base (20% of 86,463), we will retain 3,034 wagons that derailed and 14,259 wagons that did not derailed.

Thus, the same proportions are respected in both groups.

Another method consists in initially using the same number of derailed cars and non-derailed cars. To make the prediction and evaluate afterwards. Then to make the prediction with the real values as observed in the distribution. Compare the results of the two predictions; if they go in the same direction, then retain the results of the study with the complete distribution; otherwise, work by gradually adding additional individuals to the base. The addition stops when the difference in results becomes significant. Evaluate the percentage of difference, if it is large enough, the study is rejected, if not, the study is accepted (Kabeya, 2003; Kasongo, 2003). This second way of proceeding is similar, in statistics, to studying a distribution having several variables by comparing the variables two by two in order to hope to generalize the results at the end, whereas one would easily proceed by a general analysis by an appropriate method such as ANOVA (Rees, 2018). This reason made us able to reject the procedure.

For the general study, which will include images and other types of data, we will need to retain a sufficiently large and varied sample to prevent the image base from being insufficient for training. This aspect does not mean that we have to go through the entire network to identify the faults in all the places where they are, but we will retain standard images which will serve as a reference to which the other images that the drones or satellites will capture will be reported. From this comparison, a module making it possible to indicate the defects of the track could be set up, a module which will allow the teams to work before the passage of the train.

Thus done, the preparation of the data takes a lot of time. Experts in the field estimate this time at 80% of the time needed to set up a Machine Learning project (Woldeamanuel et al., 2023; Choi et al., 2021; Studer et al., 2021). This percentage shows that the quality of the models will depend on the quality of the data collected. Hence, the time spent preparing the datasets is well worth the effort.

For the generalization of the study, the data will be grouped in a Data Lake which will contain the information necessary for the processing of the fuzzy elements present. The data lake structure to be put in place must be sufficiently fed to prevent it from becoming a data swamp. For this aspect, we plan to set up a system capable of recovering the photos of the track, of relating them to the structure of the existing defects, of identifying the recognized defect(s), failing that, of updating the structure of the defects in order to constitute a solution for updating the photos. This data lake will be a new component of the system to be put in place.

Having processed and retained appropriate data, we can move on to the next step, which is determining the model and training the model by using K-NN procedure. The application of the k-NN method gave us the possibility of deciding whether a vehicle stored for departure in the line of a train will derail or not before its arrival at its destination. This decision is made in accordance with the findings made on its immediate neighbors. Thus, after application, we obtained 87% chance for the procedure with 3 neighbors and 85% with the procedure with 5 neighbors.

DISCUSSION

The goal of this study was to know what operations to apply to ensure that the data obtained is an effective set of information that can lead us to a better prediction? We have found that in any project to set up a prediction model using Machine Learning techniques, data is a very important element, often justifying the failure or success of the project.

After understanding the business, the retention and preparation of data is the second step and is by far the most important step for the realization of prediction models.

There is a saying that the predictions you have are worth the data withheld. Well-collected and well-prepared data will surely provide better predictions.

A good data preparation starts with the collection of the appropriate data; we must indicate where to find the data necessary for the study. For old companies such as the SNCC, the data is often included on handwritten sheets that we must enter or have entered. Given the amount of information required, this operation generally takes a long time. This handwritten data is often accompanied by data entered in spreadsheets taken from the operational departments that perform the tasks or from the centralization department (Planning and Internal Control Department).

Data collection is followed by determining the necessary volume of data to build a good machine learning project. These data must be numerous so that they can lead to a good prediction.

For old companies, quantity is not an issue, as enough data can easily be found for a machine learning project. The problem arises at the level of quality, the data entered manually have many errors which they are not corrected can annihilate all the expected results.

The distribution between the learning base and the test base will be made at the rate of 80% of the data for the learning base and 20% for the test base.

Pre-trained models can be applied well, but considering the specificity of the modifications made during the constitution of the training base, an appropriate model would do the trick.

Collecting and preparing the data is the most important step in the process of setting up a Machine Learning model. It collects (gathering data from different sources where it can be drawn), cleaning (“cleansing”), it is the dusting of data to make them suitable for processing by Machine Learning techniques.

CONFLICT OF INTERESTS

The authors have not declared any conflict of interests.

REFERENCES

Abaker AA, Saeed FA (2021). A comparative analysis of machine learning algorithms to build a predictive model for detecting diabetes complications. Informatica 45:1.

Adams LC (2023). Application of Data Mining and Machine Learning on Occupational Health and Safety Struck-by Incidents on South African Construction Sites: A CRISP-DM approach.

Agrawal M, Khan AU, Shukla PK (2019). Stock price prediction using technical indicators: a predictive model using optimal deep learning. Learning 6(2):7.

Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data 8:1-74.

Arnal Segura M, Bini G, Fernandez Orth D, Samaras E, Kassis M, Aisopos F, Rambla De Argila J, Paliouras G, Garrard P, Giambartolomei C, Tartaglia GG (2022). Machine learning methods applied to genotyping data capture interactions between single nucleotide variants in late onset Alzheimer's disease. Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring 14(1):e12300.

Arumugam K, Swathi Y, Sanchez DT, Mustafa M, Phoemchalard C, Phasinam K, Okoronkwo E (2022). Towards applicability of machine learning techniques in agriculture and energy sector. Materials Today: Proceedings 51:2260-2263.

Ayele WY (2021). Adapting CRISP-DM for Idea Mining: A Data Mining Process for Generating Ideas Using a Textual Dataset. arXiv preprint arXiv:2105.00574.

Beer D (2019). The social power of algorithms. The Social Power of Algorithms (pp. 1-13). Routledge.

Beheshti A, Ghodratnama S, Elahi M, Farhood H (2022). Social Data Analytics. CRC Press.

Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP (2022). Harnessing multimodal data integration to advance precision oncology. Nature Reviews Cancer 22(2):114-126.

Cai L, Zhu Y (2015). The challenges of data quality and data quality assessment in the big data era. Data Science Journal 14:2-2.

Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, Cohen Z, Belgrave D, DeRubeis R, Iniesta R, Dwyer D (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 20(2):154-170.

Chen YM, Chen PC, Lin WC, Hung KC, Chen YC, Hung CF, Wang LJ, Wu CN, Hsu CW, Kao HY (2023). Predicting new-onset post-stroke depression from real-world data using machine learning algorithm. Frontiers in Psychiatry 14:1195586.

Choi SW, Lee EB, Kim JH (2021). The engineering machine-learning automation platform (emap): A big-data-driven ai tool for contractors' sustainable management solutions for plant projects. Sustainability 13(18):10384.

Guo Q, Zheng Y, Shi J, Wang J, Li G, Li C, Fromson JA, Xu Y, Liu X, Xu H, Zhang T (2020). Immediate psychological distress in quarantined patients with COVID-19 and its association with peripheral inflammation: a mixed-method study. Brain, Behavior and Immunity 88:17-27.

Hasanzadeh S, Esmaeili B, Dodd MD (2017). Impact of construction workers' hazard identification skills on their visual attention. Journal of Construction Engineering and Management 143(10):04017070.

Heinrichs B, Eickhoff SB (2020). Your evidence? Machine learning algorithms for medical diagnosis and prediction. Human Brain Mapping 41(6) :1435-1444.

Johnstone SJ, Parrish L, Jiang H, Zhang DW, Williams V, Li S (2021). Aiding diagnosis of childhood attention-deficit/hyperactivity disorder of the inattentive presentation: Discriminant function analysis of multi-domain measures including EEG. Biological Psychology 161:108080.

Kabeya M (2003). Use of artificial intelligence in the detection of geometric defects of a railway track. The annals of the Higher Institute of Statistics of Lubumbashi (7):67-96.

Kabeya M (2021). Prediction of train derailments based on an implementation by the Machine Learning technics in railway operations. Master dissertation. Lubumbashi University.

Kamalov F, Gurrib I, Rajab K (2021). Financial forecasting with machine learning: price vs return. Kamalov, F., Gurrib, I. & Rajab, K.(2021). Financial Forecasting with Machine Learning: Price Vs Return. Journal of Computer Science 17(3):251-264.

Kasongo N (2003). Procedure for hierarchical selection of discriminate variables by the technics of step-by-step discriminant analysis in the case of several classes. Case of Male internal Medicine at Sendwe Hospital in Lubumbashi. The annals of the Higher Institute of Statistics of Lubumbashi (7):97-126

Kolyshkina I, Simoff S (2021). Interpretability of machine learning solutions in public healthcare: The CRISP-ML approach. Frontiers in Big Data 4:660206.

Kowsher M, Turaba MY, Sajed T, Rahman MM (2019). Prognosis and treatment prediction of type-2 diabetes using deep neural network and machine learning classifiers. In 2019 22nd International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

Lantz B (2019). Machine learning with R: expert techniques for predictive modeling. Packt Publishing Limited.

Madera C (2018). The evolution of information systems and architectures under the influence of big data: Data Lakes, (Doctoral dissertation. University of Montpellier), France.

Meng T, Guo X, Lian W, Deng K, Gao L, Wang Z, Huang J, Wang X, Long X, Xing B (2020). Identifying facial features and predicting patients of acromegaly using three-dimensional imaging techniques and machine learning. Frontiers in Endocrinology 11:492.

Menkel-Meadow C (2017). Portia in a different voice: speculations on a women's lawyering process. In Gender and Justice (pp. 341-365). Routledge.

Merriam SB, Grenier RS (Eds.) (2019). Qualitative research in practice: Examples for discussion and analysis. John Wiley & Sons.

Nordin N, Zainol Z, Mohd Noor MH, Lai Fong C (2021). A comparative study of machine learning techniques for suicide attempts predictive model. Health Informatics Journal 27(1):1460458221989395.

Rees DG (2018). Essential statistics. Chapman and Hall/CRC.

Rehman A, Naz S, Razzak I (2022). Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. Multimedia Systems 28(4):1339-1371.

Ribeiro R, Pilastri A, Moura C, Rodrigues F, Rocha R, Cortez P (2020). Predicting the tear strength of woven fabrics via automated machine learning: an application of the CRISP-DM methodology.

Russell B, Potter M (2022). Introduction to mathematical philosophy. Routledge.

Sarr JM, Brochier T, Brehmer P, Perrot Y, Bah A, Sarré A, Jeyid MA, Sidibeh M, El Ayoubi S (2021). Complex data labeling with deep learning methods: Lessons from fisheries acoustics. ISA Transactions 109-113-125.

Schröer C, Kruse F, Gómez JM (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science 181:526-534.

Siebert J, Joeckel L, Heidrich J, Trendowicz A, Nakamichi K, Ohashi K, Namba I, Yamamoto R, Aoyama M (2021). Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology. Machine learning and knowledge extraction 3(2):392-413.

Szepesvári C (2022). Algorithms for reinforcement learning. Springer Nature.

Toye F, Seers K, Allcock N, Briggs M, Carr E, Barker K (2014). Meta-ethnography 25 years on: challenges and insights for synthesising a large number of qualitative studies. BMC Medical Research Methodology 14(1):1-14.

Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, Cumbers S, Jonas A, McAllister KS, Myles P, Grainger D (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ, 368.

Wang R, Pan Z, Chen Y, Tan Z, Zhang J (2021). Influent Quality and Quantity Prediction in Wastewater Treatment Plant: Model Construction and Evaluation. Polish Journal of Environmental Studies 30(5).

Woldeamanuel MM, Kim T, Cho S, Kim HK (2023). Estimation of concrete strength using thermography integrated with deep-learning-based image segmentation: Case studies and economic analysis. Expert Systems with Applications 213:119249.