The Journal of Supercomputing (2021) 77:11481–11513
https://doi.org/10.1007/s11227-021-03708-5

Algorithms and software for data mining and machine learning: a critical comparative view from a systematic review of the literature

Gilda Taranto-Vera[1] · Purificación Galindo-Villardón[1] · Javier Merchán-Sánchez-Jara[1] · Julio Salazar-Pozo[1] · Alex Moreno-Salazar[1,2] · Vanessa Salazar-Villalva[1,2]

Accepted: 23 February 2021 / Published online: 25 March 2021

Abstract

Today, a greater generation of information is produced as a consequence of the technological development of society. The Internet has facilitated the access and extraction of this information, thus pursuing the automatic discovery of the knowledge contained within. In this context, data mining aims to discover patterns, profiles and trends of a large volume of data, for which multiple learning techniques are available. The selection of which technique to use depends on the type of result desired to obtain and the data that are available, considering that the algorithms for these tasks date mostly from the early twentieth century and are now the basis of these new technologies. The aim of this study is to show the development of these techniques in the field of scientific research and to present the evolution of algorithms and software for data mining in recent years. To this end, the systematic literature review methodology was applied, as it is considered a systematic process that identifies, evaluates, and interprets the work of researchers in a chosen field. As a result, we present a comparative analysis of the most outstanding software: Alteryx, TIBCO Data Science, RapidMiner and WEKA, their capacities for data mining processes and a description of the algorithms and techniques of machine learning that are currently on the rise.

Keywords Data mining · Machine learning techniques · Algorithms · Systematic literature review · Software tools · Performance evaluation

✉ Gilda Taranto-Vera
gilda.taranto@usal.es

Extended author information available on the last page of the article

1 Introduction

In recent years, the rapid progress of information technology has resulted in increased data production in almost all areas of knowledge. Also, factors such as the massive use of information on the web, the growth of wireless communications through mobile devices or the development of trends such as open data (OD) and geographic information systems have facilitated access to and extraction of these large data sets from all kinds of social and cultural activities.

All this activity is generating large volumes of unstructured data through resources such as e-mail, social networks, the production of images, text and other types of information that need to be extracted, explored, analyzed, modeled and evaluated; requiring specialized techniques that allow the identification of patterns, trends or implicit relationships.

In this context, the analysis of large data sets in different application areas allows for such significant findings as the discovery of genomic patterns, the forecasting or prediction of financial risks, the modeling or characterization of demographic groups, the automatic learning of supply chain processes or the recognition of voting preferences. This requires constant updating of computer tools and the application of the latest statistical, mathematical and computational techniques related to big data (BD) analysis, machine learning (ML), artificial intelligence (AI), data mining (DM) or business intelligence (BI), among others.

Among the previously mentioned techniques is DM a set of techniques that aim to discover patterns, profiles and trends through the analysis of data using pattern recognition technologies, neural networks, fuzzy logic, genetic algorithms and other advanced techniques of data analysis.

DM is a discipline closely related to concepts such as ML or knowledge areas such as statistics. In general, statistics is the first science that historically extracts information from data employing methodologies coming from mathematics; thus, the concept of DM arises from the application of new computer technologies to this type of task.

The ML, as a branch of AI, comprises the set of methods and techniques to detect patterns in an automated way in a data set using them to predict future data or to carry out other types of decisions in an uncertain environment [1]. Subsequently, the development of DM as an autonomous discipline has been allowed by the increase in size and structuring of the data to which these techniques are applied [2]. It is recognized as an essential tool of modern business, since its capability of converting data into BI, thus giving information advantage [3].

From its origins to the present, DM has been conceptualized from different perspectives, from its consideration as an information extraction process, from being a data warehouse (DW) modeling technique to being defined as an integral process covering from the data selection project implementation up to the follow-up within an organization.

DW as a process is contemplated from three main approaches: the KDD Methodology (Knowledge Discovery in Database) [2], the SEMMA Methodology of the SAS Institute [4] and the CRISP-DM Methodology (CRoss-Industry Standard

Process for Data Mining) [2], the latter being the most accepted in the field of scientific research due to its completeness and good documentation, where all its stages are duly organized, structured and defined, allowing a project to be easily understood or reviewed [5].

The implementation of the DM process in its modeling phase is developed through a series of ML algorithms and techniques that aim to extract the knowledge implicit in data sets from different origins. The choice of learning techniques is determined by the type of results to be obtained and the data available.

The initial classification of these techniques distinguishes between predictive, descriptive, and auxiliary techniques such as the Analytical Process of Transactions (OLAP), SQL and querying and reporting tools [2]. Besides, we can consider the technique by reinforcement oriented to "learning by themselves what to do" [6] and the semi-supervised techniques that lie between the first two [7].

It is important to emphasize each technique adequacy according to their scope of application; if the most appropriate technique is not used, there is a risk of losing all the value and richness of the information contained in the databases [8, 9].

In addition, a wide range of software has been implemented for DM and ML, which offer intelligent solutions, covering the project from problem statement to implementation and testing.

These solutions, both open source and commercial versions, have hundreds of algorithms implemented and count with the collaboration of partners for the accelerated development of more techniques.

For the categorization and evaluation of data science and ML technology platform providers, we count with the contribution of the main data science portals such as KDNuggets, Gartner and Business Over Broadway, all of them with great recognition to the scientific community.

From this perspective, the objective of the research is centered in synthesizing, in a systematized way, the evolution of the DM in recent years, to evidence the development of the techniques in the scope of the scientific research, to reach a deeper knowledge about the characteristics of the algorithms of ML, software currently most used and to identify the areas of knowledge in which they are used, in addition to their impact and scientific contribution.

For this purpose, the SLR methodology is applied, as it is considered the most powerful, formalized and rigorous type of literature review to synthesize and critically analyze multiple studies or research papers through a systematized process [10].

Once the systematic review process is concluded, a compendium of papers is analyzed and synthesized as the most relevant to understand the state of the art, around DM and ML during the chosen period.

2 Learning skills

Machine learning techniques can be defined as a set of methods capable of automatically detecting patterns in data [1], of which techniques applied to DM such as predictive or supervised learning, descriptive or unsupervised learning, reinforcement and semi-supervised techniques are being subject of analysis in this paper.

2.1 Predictive techniques or supervised learning

Supervised learning is a pattern recognition technique that allows to estimate the output of a function from a set of training data; therefore, they specify the model for the data based on previous theoretical knowledge. The model assumed for the data must be checked after the DM process before it is accepted as valid [2]. This typology of techniques is, in turn, divided into classification and regression techniques.

Classification is the process of finding a model (or function) that describes and distinguishes classes of data or concepts; this model is based on the analysis of a set of training data (i.e., the objects for which the labels are known), to predict the class of the objects whose label is unknown.

While classification predicts categorical (discrete, unordered) labels, regression models continues value functions. That is, the regression allows you to predict missing or unavailable numerical data values instead of (discrete) class labels. The term prediction refers to both numerical and class label prediction [11]. Predictive or supervised techniques include decision trees, Bayesian models, neural networks, support vector machine (SVM), linear models, instance-based learning, blended learning techniques (bagging, boosting and stacking), among others.

2.2 Descriptive techniques or unsupervised learning

Unsupervised learning techniques do not require human intervention to form a data set, previously categorized, that isto be presented to the learning algorithm. The goal of unsupervised learning is to find relevant patterns considering the distribution and composition of the data presented to it [12]. This set of techniques is divided into two large groups: segmentation or classification and association or sequence techniques.

Segmentation techniques create groups based on similarities, meaning that the groups are not previously known, but are revealed by the model. In the case of association or sequence techniques, models detect the association between discrete events, products or attributes and play the role of both predictor and target variables.

In this category are included the techniques of clustering, exploratory analysis, dimensionality reduction and multidimensional scaling, such as factorial analysis, principal component analysis, correspondence analysis [2].

2.3 Semi-supervised learning techniques

This set of techniques uses tagged and untagged data, its main challenge is to explore the information that contains the unlabeled data in an efficient way, which is more convenient, since obtaining the labeled data is generally much more expensive.

Semi-supervised learning allows to take advantage of the unlabeled data and obtain a predictive model that can work better than one that only uses labeled

data. Also, it allows to use a smaller amount of tagged data and obtain the same level of results, that is, the effort in tagging is reduced, with the consequent decrease in costs.

On the other hand, it has also been shown that semi-supervised learning techniques are not always the most efficient and productive since they depend as much as on choosing the right model to use semi-supervised learning, as depending on the quality of the unlabeled data [13].

2.4 Reinforcement learning techniques

Learning by reinforcement is a technique based on trial and error, which is used when there is no detailed information about the desired output, and unlike supervised learning, it does not have a teacher who instructs about the correct outputs in the face of certain inputs, but a critic who provides information that is more evaluative than instructional; this technique consists of mapping situations to actions, maximizing a scale called a reinforcement or reward signal [14, 15].

The elements that form a system of learning by reinforcement are an agent, the environment, a policy, a value function and, optionally, a model of the environment.

The agent is the one that learns and makes decisions; the environment is what interacts with the agent; the agent selects actions; and the environment responds to these actions by presenting new situations to the agent. From its part, the policy defines the way the agent behaves and a reward function defines the objective or goal in reinforcement learning [16].

Among the outcomess of this study, active reinforcement learning with demonstration is found to improve RL with demonstration more efficiently regarding human effort.

Under the framework, we observed a solution called active deep Q-network based on a classical RL algorithm called deep Q-network (DQN), where the algorithm dynamically estimates the uncertainty of recent states and utilizes the queried demonstration data by optimizing a supervised loss in addition to the usual DQN loss, for which two measurements of the uncertainty were provided: the divergence of bootstrapped DQN and the predictive variance of noisy DQN [17].

3 Methodology

The scientific SLR methodology carried out in the present work aims to identify, synthesize and critically analyze the most significant postulates and theoretical developments that have been developed in the chronological period between 2016 and 2020 in the field of algorithms and software for DM and ML.

Applying this type of review to emerging disciplines with a short chronological path allows not only to synthesize the most significant contributions, but also to identify gaps or important unexplored lines, and in other cases, to direct the steps towards future research [18].

The general objectives that motivate the application of an SLR are specifically aimed at identifying the most significant and relevant algorithms and software developments proposed or developed concerning the different areas of application and ultimately to offer a critical synthesis of the most relevant resources and advances that have emerged within the discipline in recent years.

The methodology used to articulate the present SLR follows, in general terms, the sequence of processes proposed in [18] and contemplates the following stages:

i Definition of the research question.
ii Delimitation of scope and objectives.
iii Identification of databases and sources.
iv Definition of scope of action:Population, Intervention, Comparison, Outcomes, Context (PICOC).
v Definition of inclusion and exclusion criteria Formulation and execution of search equation.
vi Selection of papers using title and abstract or text exploration.
vii Quality control by applying the inclusion and exclusion criteria.
viii Forming of corpus of papers.
ix Analysis and extraction of information.
x Synthesis and elaboration of the report.

3.1 Definition of the research question

The research questions that guided the review process were:

RQ1: What are the most used DM techniques?
RQ2: What are the most prominent ML algorithms today?
RQ3: In what areas of knowledge are the DM techniques applied?
RQ4: What are the characteristics and capabilities of the main DM software?
RQ5: How many studies have been published between 2016 and 2020 in the field of DM? (Fig. 1).

3.2 Identification and selection of sources and databases

After an exploratory process aimed at mapping the particularities of the scientific information communication practices in this discipline, which are certainly new, it was decided to group the sources, at a generic level, into three main groups: databases, repositories and web resources, digital scientific journals (commercial and open access) and academic monographs.

The scope of action of the present SLR is contextualized through the application of the following issues. The criteria derived from them were applied in the documentary analysis to effectively answer the research questions posed.

-Population (P). Algorithms and software or applications or methods for the DM AND

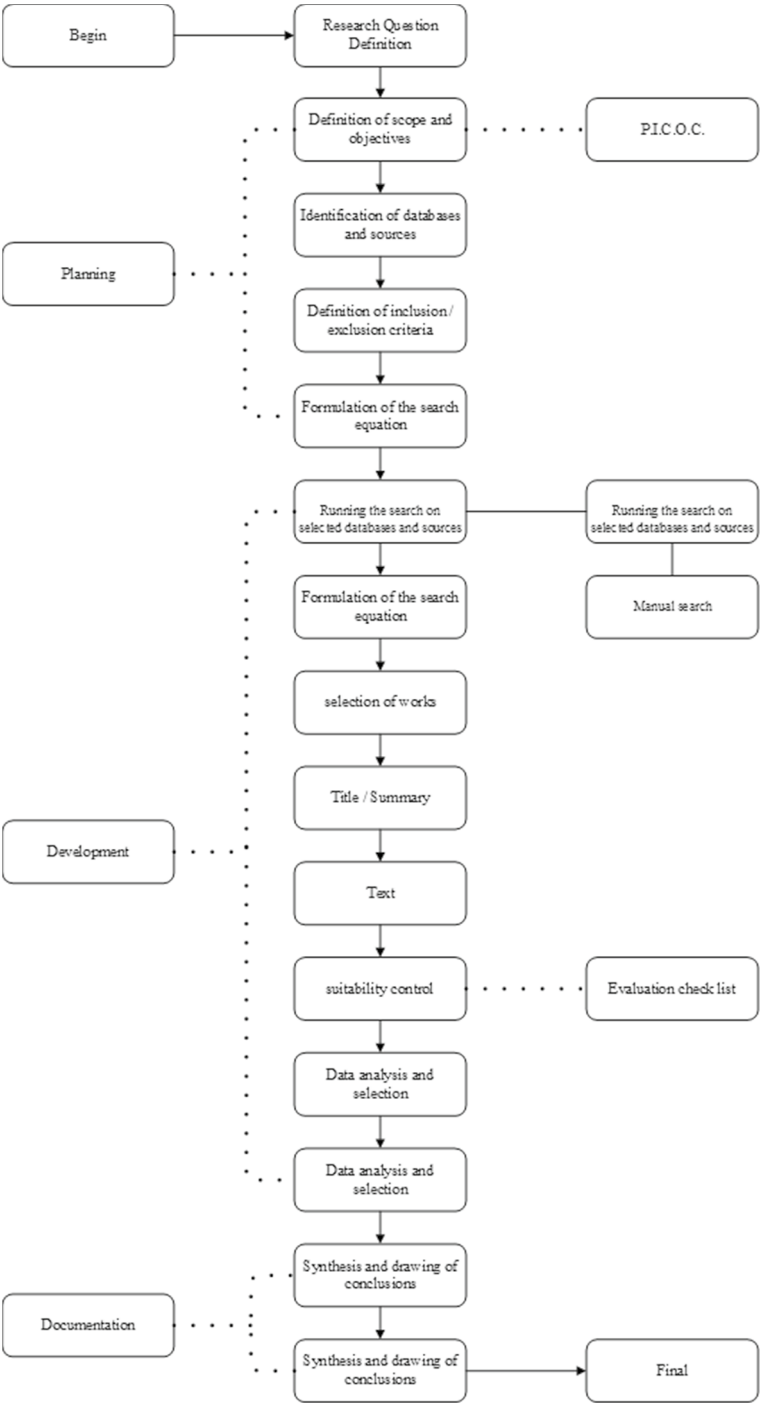

Fig. 1 Flow of methodology used to articulate this SLR

-Intervention (I).	Software proposals with ML algorithms for DM analysis in the period 2016-2020 AND
-Comparison (C).	With other literature reviews compiling studies concerning software with ML algorithms used for DM, applied from 2016 to 2020 AND
-Outcomes (O).	Outstanding software with implementation of ML algorithms for the analysis of DM at present AND
-Context (C).	Software that uses prediction, regression, clustering, association, and ML algorithms for DM.

3.3 Search strategies and formulation

Prior to the confirmation of the relevant and effective search equation, a series of exploratory searches of generic character were carried out in the Scopus and Google Scholar database, to establish the basic semantic and terminological field, by which the most relevant works in these repositories were indexed.

From these terms identification, the first searches were launched in the Web of Science, Scopus, IEEExplorer, Mendeley and Scielo databases, to obtain the most exhaustive and relevant results possible. To achieve this goal, both the search equations and the search fields in the databases were systematically refined.

Before launching the final searches and the creation of the relevant alerts, the search equations were subjected to a pretest to check that most of the references retrieved corresponded to the subject matter and objectives sought. Ultimately, this process was aimed at avoiding common errors caused by polysemy and homonymy issues in the selected search terms.

The basic semantic and terminological field was made up of the following terms:

Datamining, Data mining/Algorithms, Algorithm, algorithmic/Machine Learning/Software, applications, programs, systems /Statistical Analysis, statistics, statistical/Supervised/Unsupervised/Predictor, predictive/Descriptive/Cluster, clustering/Top ten, ranking/Review, overview.

In the generalist databases (Scopus and Web of Science), the question was posed, and the corresponding alert was created, aiming at recovering references, see Table 1:

A second search is carried out for the recovery of results in search engines and specific bibliographic databases such as IEEExplorer, Scielo, and Mendeley for their contribution to the scientific community and for being involved in this study; as well as the main portals for Data Science KDNuggets, Gartner and Business Over Broadway.

To refine the search, the following specific inclusion and exclusion criteria were created.

Inclusion:

IC1	Documents published from 2016 to 2020 AND
IC2	The publication is in English or Spanish language AND
IC3	The research shows a ranking or the top ten algorithms that apply DM.

IC4 The applied open-source and commercial software.
IC5 The applied improved ML algorithms.
IC6 The classification of the techniques used in DM.

Exclusion:

EC1 Documents published outside the year range from 2016 to 2020 AND
EC2 Publication is in a language other than English or Spanish AND
EC3 Pilot software that have, not been implemented in real cases AND
EC4 Algorithms that have not been implemented in any official computer system, package or software library (Fig. 2).

Summary of results:

Total jobs recovered:	1181
Total work selected after another application of criteria:	855
Total work analyzed:	19

4 Results analysis

Once the systematic review process is concluded, a compendium of 19 papers is analyzed and synthesized as the most relevant to understand the state of the art and the most significant developments around DM and ML during the period 2016–2020; furthermore, the most relevant techniques and algorithms are identified, as well as the most outstanding software according to scientific literature in the field of data science in 2020 (Table 2).

Within the analysis, it is appreciated that the Clustering and Classification techniques stood out among the studied topics. From group,it is distinguished **(a) evolutionary algorithms (EA)**

The EAs, which although are not learning techniques [19], haves also been applied for learning and knowledge extraction and are widely applied to complement or even replace the classical DM learning approaches, as long as it is directly motivated by the nature of the problem, resulting in the adoption of the term EDM (evolutional data mining) [20].

On the other hand, framed in the category of evolutionary algorithms (EA) are the genetic algorithms (GA) which are optimization, search and learning algorithms, name inspired by biological evolution and its genetic-molecular basis.

Many competent pattern mining algorithms have been developed in the last two decades; however, when applied to big data, they suffer an extreme computational cost in the search for association rules, being the GA a solution to optimize association rules generated by other algorithms, a challenge that includes capture, storage, search, transfer, analysis and visualization of large data sets [21].

Table 1 The basic semantic and terminological field. *Source*: self-made

Database	Search string
SCOPUS	(TITLE-ABS-KEY(DataMining OR Data Mining)) AND (TITLE-ABS- KEY("Algorithm supervised" OR "Algorithm Unsupervised")) AND (TITLE-ABS-KEY (Software OR Program OR Platform OR Application)) AND (TITLE-ABS-KEY(Supervised OR unsupervised OR predictive OR descriptive OR "statistical methods" OR "predictable models" OR predict)) AND (LIMIT-TO (PUBYEAR,2020) OR LIMIT-TO (PUBYEAR,2019) OR LIMIT-TO (PUBYEAR,2018) OR LIMIT-TO (PUBYEAR,2017) OR LIMIT-TO (PUBYEAR,2016)) AND (LIMIT-TO (SUBJAREA,"COMP") OR LIMIT-TO (SUBJAREA,"MATH") OR LIMIT-TO (SUBJAREA,"DECI") OR LIMIT-TO (SUBJAREA,"SOCI") OR LIMIT-TO (SUBJAREA,"BUSI")) AND (LIMIT- TO (LANGUAGE,"English") OR LIMIT-TO (LANGUAGE,"Spanish")) AND (LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO (DOCTY- PE,"ar") OR LIMIT-TO (DOCTYPE,"ch")) AND (LIMIT-TO (SUBJAREA,"ENGI") OR LIMIT-TO (SUBJAREA,"MATE"))

Table 1 (continued)

Database	Search string
WOS	((TS=(”Software*” OR ”Program*” OR ”Platform*” OR ”Application*”) AND TS=(”Algorithm* supervised*” OR ”Algorithm* Unsupervised*”) OR TI=(”Algorithm* supervised*” OR”Algorithm* Unsupervised*”) AND TS=(”predictive*” OR ”descriptive*” OR ”statistical methods*” OR ”predictable* models*” OR ”predict*”) OR TI=(”predictive*” OR ”descriptive*” OR ”statistical methods*” OR ”predictable* models*” OR ”predict*”) AND TS=(”DataMining*” OR ”Data Mining*”) OR TI=(”DataMining*” OR ”Data Mining*”)) AND SU=(”cp*” OR ”ar*” OR ”ch*” OR ”ENGI*” OR ”MATE*”)) and (LANGUAGE (English) AND (Spanish) AND PERIOD (2016 - 2020))

(b) Clustering The following algorithms were distinguished in this group, see Table 3.

As an example, the following are described as the two most significant.

IDBSCAN Algorithm for Spatial Data Mining. There are several DM grouping techniques for tasks as complex as knowledge extraction from a large amount of spatial data, collected from various applications such as geographic information system (GIS), satellite images, X-ray crystallography, environmental assessment and planning, air traffic controllers, etc.

One of the best-known algorithms to solve these tasks is DBSCAN (density-based spatial clustering of noisy applications), a pioneering density-based algorithm; however, DBSCAN may present some problems and requirements: a) It requires user input to specify parameter values to be executed, b) it is prone to a dilemma when deciding significant clusters from data sets with varying densities, and c) it incurs some computational complexity.

From this point of view, many researchers have tried to improve the basic DBSCAN algorithm to overcome these drawbacks [22] by developing evolutions and modifications such as VDBSCAN, FDBSCAN, DD DBSCAN and IDBSCAN.

Among the methods and algorithms mentioned above is IDBSCAN, which demonstrates multiple advantages over DBSCAN, in terms of greater accuracy in the matter of density accessibility, better performance for near border points, detection of outliers, requiring 5 parameters as opposed to DBSCAN that requires 2 parameters, groups spatial data of any type as .png, .dbf, .csv, .rgs, .xls, .scr and .txt.

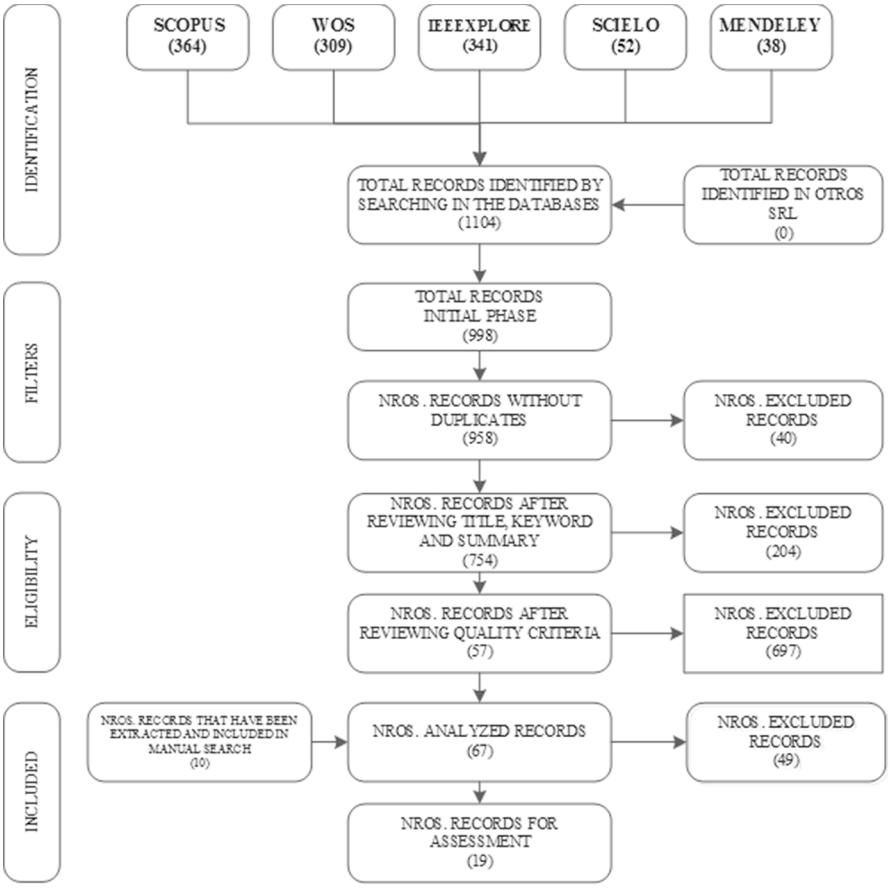

Fig. 2 Resulting SLR flow

And although both algorithms do not require the number of groups (κ) in advance, DBSCAN cannot group well the data sets with large differences in densities, while IDBSCAN can do it correctly.

The effectiveness of the proposed algorithm was demonstrated by using a real and synthetic database, providing the opportunity to apply a clustering algorithm on new data types and new application areas, such as moving objects and trajectories, spatially embedded social networks and geocoded and web-based multimedia data [23].

OMRF Algorithm for Change Detection and Image Grouping Image shift detection is a process that identifies the difference between two multispectral images acquired in the same geographic area at different times, changes can often occur due to seasonal factors such as deforestation, natural disturbances, etc. [24].

This field is evolving very quickly and constantly, finding new proposals such as a new algorithm for detecting unsupervised changes for remote sensing images

Table 2 Relevant publications about Evolutionary/generic algorithms. *Source*: self-made

Model type: evolutionary/generic algorithms			
Year	Authors	Title	DM problem
2016	García et al.	SubGroup Discovery With Evolutionary Fuzzy System in R the Sdesr Package [50]	Fuzzy Data Clustering
2017	Deshpande et al.	Expert System For Retrieval of documents using evolutionary approaches Incorporating clustering [51]	Clustering
2018	Alcalá et al.	Evolutionary Data Mining and Applications a revision on the Most Cited papers from the last [20]	Classification, regression and clustering
2018	Babi et al.	Mining Frequent Patterns from Big Data sets using genetic algorithm [21]	Association Rules

Table 3 Relevant publications about clustering. *Source*: self-made

DM problem: clustering			
Year	Authors	Title	Model type
2016	Sharma et al.	Improved Density-Based Spatial Clustering of Applications of Noise Clustering Algorithm for Knowledge Discovery in Spatial Data [23]	Spatial Argument Based on the Improved Density of the Noise Application (DBSCAN)
2016	Li et al.	Unsupervised Change Detection for Remote Sensing Images Based on Object-Based MRF and Stacked Autoencoders [24]	Unsupervised Change Detection Algorithm for Remote Detection Images Based on MRF objects (OMRF) and stacked autoencoders (SAE)
2017	Shi and Du	Manifold Regularized Robust Unsupervised Feature Selection for Image Clustering [52]	Manifold Spanning Tree-Based Clustering
2018	Haque and Minero	Contextual Outlier Detection in Sensor Data Using Minimum Spanning Tree-Based Clustering [26]	Minimum Spanning Tree-Based Clustering
2019	Sarala and Saravanan	Selection and Descriptive Self-Organized Map Clustering for Spatio-Temporal Pattern Discovery	Heuristic algorithm Best-First Search HBFS-DSOFM [53]

based on object-based MRF (OMRF) and stacked autoencoders (SAE); where a network of SAE is presented to create a detector that can learn how to analyze the images to be detected and recognize changed pixels and unchanged pixels, with the reference of predefined images just obtained by the object-based MRF model.

According to the researchers, the results of the experiment show that the overall error rate is reduced, and the accuracy of change detection is promoted. They conclude that SAE plays a substantial role in improving the effectiveness of change detection due to its powerful feature extraction capability [24].

Another effective approach to finding changes experienced over a period of time is proposed through a multiple dilated convolutional neural network, for the detection of changed and unmodified areas. The feature map determined in the fully connected network extracts the global information by expanding the receptive field covering all regions with no missing parts in the image. Therefore, this type of architecture improves learning in deep networks by detecting changes accurately.

Finally, the trained network model achieves a feature map of the two images with the result classified as modified and unmodified pixels. The accuracy of the proposed change detection result provides better results compared to existing algorithms [25] (Tables 4, 5).

(c) Fuzzy Data Clustering/(d) Anomaly Detection

To illustrate, the most significant finding of these groups (c) and (d) is described below.

Detection of outliers. Algorithm: Teda-Cloud

Outlier detection is a fundamental data science task with a wide range of applications, including fraud detection, network security, environmental monitoring, agricultural management and public health surveillance.

The detection of outliers in sequential data has been extensively studied in the DM field and several techniques have been developed to perform this task, most of which are monitored in nature and therefore require tagged data to train the model [26].

In this context, an algorithm called Teda-Cloud has been developed, based on the recently introduced Teda (Typicality and Eccentricity Data Analytics) approach. Teda-Cloud is a statistical method based on the concepts of typicality and eccentricity that allow the grouping of similar data observations.

Instead of the traditional concept of clusters, data are grouped in the form of granular units called data clouds, which constitute a type of structure without a predefined form or established limits. Teda-Cloud is an algorithm under development, which can be used for the grouping of online data flows and applications that require real-time response. Since it is autonomous, Teda-Cloud can ”start from scratch” (from an empty knowledge base), create, update and merge data clouds, without requiring any user-defined parameters (cluster number, size, radius) or previous training.

Besides, Teda-Cloud does not rely on a specific data distribution or assumption of independence of data samples unlike most traditional statistical approaches. It is known from the literature that the results obtained from multiple data sets are very encouraging [27].

Table 4 Relevant publications about the fuzzy data clustering. *Source*: self-made

Year	Authors	Title	DM problem	Model type
2018	Costa et al.	A New Evolving Clustering Algorithm for online Data Streams [27]	Fuzzy Data Clustering	Teda-Cloud
2019	Suo et al.	Application of Clustering Analysis in Brain Gene Data Based On Deep Learning [54]	Fuzzy Data Clustering/Deep Learning	Fuzzy C-Means (FCM) y Deep Belief Network (Dbn)

Table 5 Relevant publications about the classification and anomaly detection. *Source*: self-made

Year	Authors	Title	DM problem	Model type
2018	Sumalatha and Santhi	A Machine Learning-Driven Approach to improve efficiency of Classification Algorithm using Prediction of Affliction [55]	Prediction/Classification	Model the Classification CAPM
2018	Femi and Ganesh	Comparative Study of Outlier Detection approaches [56]	Classification, Regression, Segmentation and Clustering	Anomaly/Deviation Detection

There are also other proposals for detecting contextual outliers in sensor data using clustering techniques based on minimal expansion trees. 2018 IEEE [26].

e) Text Classification Method/Sentiment Analysis

In recent years, the explosive growth of online media, such as blogs and social networking sites, has allowed individuals and organizations to write about their personal experiences and express their opinions (Table 6).

In this case, one of the algorithms to highlight is the SA SYSTEMS in which text mining seeks to predict these feelings (online text messages, tweets, reviews, etc.) through a text classification method based on unsupervised dependency analysis. This algorithm aims to predict whether an online text expresses positive, negative, or neutral feelings without the need for supervision, unlike other existing learning or lexicon-based systems that only take into account isolated words and not the relationships between them.

Sa systems detect the feeling in both short messages (tweets and SMS) and reviews. The focus is on determining the dependencies between lemmatized tagged words using a Sentiment propagation algorithm that considers and distinguishes between key linguistic phenomena, intensification, modification, denial and adverse and concessional relationships.

This proposal consists of four stages: pre-processing, lexical and syntactic analysis, creation of a sentiment lexicon and sentiment analysis through propagation [28].

Sequential Pattern Discovery using Equivalent Class (CSPADE) and Solving Constraint Integer Programs (SCIP)

There are also other tasks inherent in the text classification method where large volumes of online data are generated, which are the massive and open online courses (MOOCs) that have experienced rapid development; however, one of the main problems of online education is the high dropout rates of participants.

Many studies have used both quantitative and qualitative methods for this analysis, in an attempt to explore this problem; however, there is a lack of studies that predict the actual time of dropout, providing opportunities to improve MOOCs student retention by offering timely interventions.

For these tasks, we proposed to analyze some methods, among them Search for Explanations of Cluster of Process Instances (SECPI), a valuable technique for analyzing behavioral data that explains and discriminates whether an individual is in the dropout or non-dropout group; however, SECPI only provides rules at the instance level and, therefore, would not be useful for dropout prediction, unless individual rules are aggregated and general dropout patterns can be identified, while CSPADE and SCIP provide a result that is maintained at the group level and can be configured into an actual dropout prediction operationalization.

Among other things, this method aims to predict student dropout using process and sequence mining techniques. While process mining is capable of descriptive analysis, sequence mining techniques offer better characteristics for predictive purposes. Springer International Publishing AG 2018 [29].

Others results

Besides, from the analysis of the 19 resulting publications, information was obtained from one of the most influential and widely accepted sources in the data science community, the KDNuggets portal (Knowledge Discovery Nuggets www.

Table 6 Relevant publications about the text classification method. *Source*: self-made

Model type: sentiment analysis			
Year	Authors	Title	DM problem
2016	Fernández-Gavilanes et al.	Unsupervised Method for Sentiment Analysis in Online texts [28]	Texts C M
2018	Deeva et al.	Dropout Prediction in MOOC A Comparison between Process and Sequence Mining [29]	
2018	Li	Feature Extraction and Learning Effect Analysis for MOOCs users Based on Data Mining [57]	
2019	Abd El-Jawad et al.	Sentiment Analysis of Social Media Networks Using Machine Learning [58]	Text Classification Method/Deep Learning

kdnuggets. com/@kdnuggets, which is taken as a reference in the articles and websites about data science, which have been part of this research. According to the community survey that KDNuggets carries out annually, where the main methods and algorithms of data science/machine learning used in the period 2018 - 2020 for a real-world application are addressed, there is an increase in the use of neural network techniques and deep learning, while the decomposition of singular value decomposition (SVD), support vector machine (SVM) and association rules, among others, show a decrease.

Below are the ML-related algorithms that have experienced the greatest growth, see Table 7.

For the analysis and better understanding of the algorithms mentioned in Table 7, we refer below to the concepts of neural networks (NN) and its extensions:

NNs are one of the classification tools used in soft computing techniques that are based on the mimicry of the human brain, under the concept of a network of artificial interconnected nodes to process and provide information.

An extension of the NN is, Deep learning (DL), an emerging field, which with rapid and remarkable growth for more than ten years, is gaining more and more attention from different researchers. DL which is a branch of ML, consists of a probability system that allows computational models composed of multiple processing layers to learn from data with multiple levels of abstraction [30].

Compared to shallow architectures, it has great advantages in both feature extraction and model fitting. Also, it is very powerful in discovering increasingly abstract feature representations whose generalization ability is very strong from raw input data; it has also successfully solved some problems that were considered difficult to solve in the past in AI.

Besides, with the notable increase in the size of data used for training and dramatic increases in chip processing capabilities, NN has resulted in significant progress and has been used in a wide area of applications such as object detection, computer vision, natural language processing, voice recognition and semantic analysis [31].

The DL algorithms on the rise are described below:

Convolutional Neural Networks (CNN)

CNNs are multi-layered hierarchical NNs, whose convolutional layers alternate with the subsampling layers, reminiscent of simple and complex cells in the primary visual cortex HUBEL and WIESEL [32]. CNNs vary in how convolutional and subsampling layers are made and how networks are trained.

Table 7 List of ascending algorithms for the year 2019. *Source*: KDNuggets Portal

Algorithm	%
Generative adversarial networks (GAN)	↑101.8
Recurrent neural networks (RNN)	↑56.5
Convolutional neural network (CNN)	↑37.5
Ensemble methods boosting	↑35.5

Among the main strengths of CNN are video recognition, intelligent surveillance, multimedia understanding, image classification, facial recognition, audio retrieval, electrocardiogram classification, object detection and other fields. The central problem of CNN's extension from image to video is the exploitation of temporal information because CNN is designed primarily to extract 2D spatial features from still images and videos are naturally viewed as 3D spatial-temporal signals [33].

Computational speed is also a limiting factor for CNN architectures characterized by many building blocks typically established by trial and error [32].

The most popular CNNs for object detection and object category classification from images are Alex Nets, GoogLeNet, and ResNet50; GoogLeNet and ResNet50 being able to recognize objects with better accuracy compared to Alex Nets [34].

In fields such as medicine, the application of a CNN allows modifying existing models to decrease the cost in time; another factor is to make the best use of a pre-trained CNN by learning transfer and fine-tuning and also to take advantage of CNN models for feature extraction and differentiation of malignant lesions from benign ones, through the use of automatic learning classifiers [35].

Recurrent Neural Networks (RNN) Countless learning tasks require dealing with sequential data. Image captioning, voice synthesis and music generation require a model to produce such results. In other domains, such as time series prediction, video analysis, and music information retrieval, a model must learn from the inputs that are sequence. Interactive tasks, such as translating a natural language, engaging in dialogue and controlling a robot, often require both capabilities.

Recurrent neural networks (RNN) are connectionist models that capture the dynamics of sequences through cycles in the network of nodes. Although RNNs have traditionally been difficult to train and often contain millions of parameters, recent advances in network architectures, optimization techniques, and parallel computing have enabled successful learning in this area.

Therefore, in recent years, systems based on long-term memory (LSTM) and two-way architectures (BRNN) have demonstrated innovative performance in tasks as varied as image subtitling, language translation and handwriting recognition [36].

Generative adversarial networks (GAN) The exponential development of AI research has enabled a very significant advance in the ability of machines to imitate human nature. The generative adversarial networks (GANs) algorithm that was introduced by Goodfellow et al. [31] is inspired by the zero-sum game for two players, in which the total winnings of two players are zero, and the profit or loss of each player is exactly balanced by the profit or loss of another player [37].

GAN is composed of two models: the first is a generative model (G), which attempts to create new instances that are completely different from those of the real data set and that were never seen by the generator but have a data distribution, similar to the real data set one. The generator receives a random noise, which it uses to create new samples that obey the real data distribution. The second model is a discriminator model (D); job is to look at the distribution that could have come from the original data set or through the generator (G), and to estimate the probability that the sample is from the original data set or from the generator; the second model learns from the discriminator's comments since it does not have access to the actual data set.

On the other hand, the discriminator has access to both the real data samples and the samples taken by the generator (dummy samples). The error for the discriminator is provided by how many times, the discriminator classifies a false synthesized sample, from the generator to the real data set.

The same may be a measure of error for the generator, for example, how many times the false sample was classified as false by the discriminator [38].

Recently, Hidasi et al. [39] apply recurrent neural networks (RNN) with gated recurrent units (GRU) for session-based recommendation. The model considers the first item clicked by a user as the initial input of RNN, and generates recommendations based on it. Then the user might click one of the recommendations, which is fed into RNN next, and the successive recommendations are produced based on the whole previous clicks.

Tan et al. [40] further improve this RNN-based model by utilizing two crucial techniques, however, some problems regarding the effectiveness of the session-based recommendation method remain open, this greatly limits their application in real-world scenarios To improve the recommendation performance in the news domain and to address the session-based recommendation problems, a novel dynamic attention integrated neural network (DAINN) was proposed. The basic idea of this model is to build a unified representation of the current user, and then generate predictions on the user's next possible event with it. The representation should consider various potential factors that influence the user's next decision.

Later, to improve the recommendation accuracy, dynamic topic modeling [41] and convolutional neural network (CNN) sentence model (Kim 2014) was adopted to effectively learn the item semantic embedding. More importantly, to handle diverse variance of users' clicking behavior was introduced a novel attention scheme that would dynamically assign influence factors on recent models based on the users' spatio-temporal reading characteristics [17].

It is important to highlight among the findings of this research, the combination of classical and ML techniques. In this case, data mining technology is applied to solve the problem of the expected return of currency in the blockchain.

The support vector machine (SVM) and the semantic orientation pointwise mutual information (SO-PMI) algorithms are combined to realize the judgment of the price of the newly issued currency in initial coin offerings (ICO) blockchain, and then the corresponding prediction is made according to the trend to realize the analysis of the expected return of a new currency in the blockchain.

In the process of data mining, the combination of the two algorithms can accurately analyze the collected resources and make accurate judgments.

The strategy is to extract the communication information of the global mainstream-encrypted currency community and relevant blockchain professional's social platform release information.

Then, based on the emotional analysis of users, the return of money is predicted. In the SO-PMI algorithm of data mining, the TRIE index tree and its segmentation dictionary mechanism serve for the analysis of users' emotions.

Therefore, first, how to build user sentiment analysis based on the SO-PMI algorithm is explained, and then the benefits of a blockchain new currency are predicted [42].

Leading software for DM Another topic of our research was the software for DM and ML that stands out in the middle, using the Gartner portal (https://www.gartner.com), which is attributed the so-called magic quadrant (MQ) for the categorization of the evaluation of the providers of technological platforms of data science and ML used by data scientists.

We made a comparison between the results of 2016 and 2018, there being a notable difference between both years framed mainly by the addition of ML to the name of the Gartner report and in its title the terms ”data science” and ”machine learning”, called until the year 2017 ”Magic Quadrant for advanced data analysis platforms”. The new name of the MQ reflects the ML, the momentum, and its important contribution to the broader discipline of data science.

The axes of the MQ categorize the participants according to their ability or capacity of execution and capacity of vision, placing them in the categories: Leaders, Challengers, Visionaries and Niches players, as appropriate.

Below is a comparison between the results published by Gartner in 2019 and 2020, with the behavior of the most prominent software (Fig. 3).

It is worth mentioning that the period analyzed in the MQ corresponds to November 2018 and 2019.

Based on the results of the Gartner MQ for data science and machine learning platforms, we proceed to deepen the analysis of some of the group’s tools; this comparison is made: Alteryx, RapidMiner and TIBCO Data Science, as they consider their performance and evolution relevant, and, the special case of WEKA, which, as it is not commercial, is not found in MQ, nevertheless, it is one of the most complete open code platforms and co-participates in the development of the techniques implemented in other software.

It is also worth mentioning that in 2005 the Association for Computing Machinery awarded WEKA with the ”ySIGKDD Service Award” for its high contribution to research. This software is the basis of the ML reference work first published in 1999 by Eibe Frank and Ian H. Witten entitled ”Practical Machine Learning Tools and Techniques”. Compared to other DM tools, WEKA has proven to be particularly useful in the field of teaching and research.

The following is a brief description of each software:

Alteryx It is a registered trademark of Alteryx, Inc. offering an end-to-end analytics platform that empowers data analysts and scientists alike to break down data barriers and provide information, demonstrating in this latest MQ report, a strong focus, especially on process automation, through the ML and AI.

The features engineering capabilities combined with the platform’s ease-of-use create an environment that further accelerates time to understanding and time to value. In addition to developing an innovative set of algorithms to optimize the slow, error-prone manual process, it automates the model fitting process, hyperparameter fitting, feature engineering and data preparation process, resulting in significant increases in model accuracy and overall process efficiency.

The main tools of Alteryx allow: Better data understanding through holistic data profiling, as well as trend and outlier detection and global problem solving with language switching, using Alteryx Designer.

Fig. 3 Gartner Magic Quadrant Chart Source: KDNuggets Portal (2019–2020)

Manage users more effectively on Alteryx Server with simplified role management so you can make massive permission changes and see all users and roles in a new, unified design. Enrich analysis and find the best location for a business through its updated TrafficMetrix traffic data counting system Access Virtual Folders in Alteryx Connect by providing saved custom views.

Leverage the power of Mongo Atlas with TLS/SSL connections to support your business strategy and navigate Alteryx Server in the most secure and convenient way.

Improve data transfer from Alteryx to Snowflake with the enhanced Snowflake Bulk Loader, eliminating additional costs and processes, ensuring data is instantly available and stored faster.

Another aspect by which Alteryx maintains its leadership in the market is by continuing to develop its best platform, without code, friendly and contributing to the community of open-source data scientists by allowing to take advantage of Feature Lab open-source libraries with more than 350,000 downloads.

TIBCO Spotfire Data Science With its data visualization, advanced analytics and data management platforms, TIBCO Spotfire Data Science and business analytics platform enables data scientists and business users to collaborate on advanced analysis through massive, scalable databases processing [43].

TIBCO has a remarkable ”Vision Capability” according to Gartner because of a robust and growing product portfolio that houses all the company’s data and advanced analytical functionality (Spotfire /TIBCO Data Science Statistics and

StreamBase). TIBCO's Data Science offering is comprehensive and easy to use, featuring an impressive set of connectors and APIs for machine data capture and model scoring, data management (data virtualization and DMM), visualization capabilities (Spotfire and Jaspersoft), robust descriptive analytics and visionary predictive analytics.

According to KDNuggets, he says that TIBCO "In addition to having end-to-end development and implementation capabilities, TIBCO successfully addresses the under-served data science Internet of Thing (IoT) analysis domain. Visual workflows that span the entire data science process, from data ingestion to model management, provide a solid foundation for effective collaboration of all roles (data scientists, business analysts, citizen data scientists, process engineers, etc.) [44].

RapidMiner It is a DM tool created by Yale University in the United States, developed in the Java language; in the year 2020, it had a significant change when it descended from the Leader group to the Visionaries group, attributed to its slower growth about other suppliers in this MQ.

This software allows the development of data analysis processes by chaining 500 operators through a graphical environment, allows the use of the algorithms included in WEKA and contains data pre-processing techniques, predictive and descriptive modeling, training methods and model testing, data visualization, automatic learning [45].

RapidMiner is involved in every step of the DM process and is also involved in the visualization of the results. The tool is made up of three large modules: RapidMiner Studio, RapidMiner Server and RapidMiner Radoop, each responsible for a different DM technique. It also prepares data before analysis and optimizes it for rapid processing. For each of these three modules, there is a free and a commercial version.

The strength of RapidMiner, compared to other DM software, lies in predictive analysis, i.e., the forecasting of future developments based on the data collected [46].

WEKA Acronym for Waikato Environment for Knowledge Analysis, it is an environment for data analysis experimentation that allows the application, analysis, and evaluation of the most relevant data analysis techniques, mainly those coming from the ML, on any user data set. This tool for the HLS and DM designed in Java is distributed under the GNU-GLP license contains a collection of algorithms for data analysis and predictive modeling, allows data visualization, and provides a graphical interface [47].

WEKA is open source and can be downloaded from the University of Waikato website at http://www.cs.waikato.ac.nz/ml/weka/. It can be consulted from the reference manual for the application and other related publications, such as downloading examples for testing with this tool.

This program has three graphic working environments and a console mode environment, allowing the implementation of algorithms for data pre-processing, classification, clustering, attribute selection, association rules [48]. Currently, the stable version 3.8.3 has been released (it receives only bug fixes) and the version in development 3.9.3 (receives new features).

5 Comparative analysis

The general and special features of the software under study are described below (Table 8, 9):

Also, these platforms have equivalent tools that allow data conversion, data processing filters, model validation, model combination, data visualization, ease of configuration and ease of installation as a discrete element WEKA does not have a user-friendly interface.

5.1 Capabilities to support the DM process

The CRISP-DM methodology has been designed to support the DM process.

For this, it is interesting to mention the platforms and products that this software have, such as:

In terms of **data preparation**, WEKA, RapidMiner, TIBCO Data Science and Alteryx have tools that allow them to work with structured, unstructured and semi-structured data, as well as variable selection, data integration, projection, cleaning and transformation, but not with the automated assistance for data preparation that only TIBCO Spotfire Data Science has.

According to the **data modeling** phase, the four software packages studied can perform descriptive statistics, supervised, unsupervised, semi-supervised algorithms, by reinforcement, blended learning, automated assistance for data modeling and validation.

5.2 Fields of application of DM and ML techniques

Today, the DM has evolved in almost any type of context, even organizations have been forced to change business strategies to accommodate decision making based on information analysis and knowledge extraction from the large volumes of data (big data) they handle [49]; as an example, we can mention (Table 12):

6 Discussion and conclusions

Considering the wide range of topics addressed in this research and because it was evident that there is no updated systematic review that meets the criteria and objectives of this study, we proceeded to apply a methodology such as the SLR, which allowed us to collect and analyze a large number of publications; thus, we have attempted to provide a valuable resource for those seeking a current view of DM, ML algorithms, the most prominent software in this field and areas of application.

In this context, it is important to mention that there are several concepts about the DM process that differ and there are methodologies that consider it only as a

Table 8 Comparison of general characteristics of the relevant software. *Source*: Self-made

Characteristic	Software			
	WEKA	RapidMiner	TIBCO Data Science	Alteryx
Licencia	GLP	AGPL 3.0	EULA (End User License Agreement)	EULA (End User License Agreement)
Basic Version	–	RapidMiner Studio Community	TIBCO Clarity	Alteryx Analytics
Professional Version	WEKA 3.9.4	RapidMiner Studio Professional	TIBCO Data Science 6.5.0	Altery Designer
Programming Language	Java			
operating system	Multiplatform: Windows, MacOs, Linux			
Other versions	WEKA Server: ADes, Aut	RapidMiner Server, Col, ADes, Desp, RapidMiner Cloud, Cloud repository to store projects	Spotfire Server: Rest and SOAP API's	Alteryx Server: REST API estandar, JAVA, Python, Php, Saleforce, Node.js
BigData (+1 millon GB)	Distributed WEKA Hadoop/Spark	RapidMiner Radoop	TIBCO Spotfire Visual Analytics for Hadoop/Apache Hive, Apache Spark SQL, Cloudera Hive, Cloudera Impala, otros	Alteryx Designer
TextMining	Yes (Classification Algorithms + popular J48)	Text Mining	TIBCO Spotfire Data Catalog Partner Attivio	RegEx
Data Stream	Massive Online Analysis (MDA)	Data Stream	TIBCO Streaming	Alteryx SDK API
Web Mining	Yes	Semantic Web Mining	TIBCO Streaming	
Image Mining	Si	Image Processing	TIBCO Spotfire (con IronPython, AWS)	Alteryx SDK API
Geospatial Analysis	WEKA GDPM	RapidMiner Studio Core (DBSCAN)	TIBCO Spotfire GeoAnalytic	Alteryx Designer Location Intelligence and GeoSpatial Analytics

Table 9 Availability of software modeling algorithms. *Source*: self made

Algorithm type	WEKA	Rapid miner	TIBCO Spotfire	Alteryx
Supervised	111	84	39	46
Unsupervisado	18	68	7	19
Extended	40+ (Pentaho Community)	100+ (Rapid Miner Community originarios WEKA)	+ Algorithm R, Python	30 Algorithm R, Python
Total	169+	252+	46+	65+

modeling technique from a DW (data warehouse); however, the most accepted concept in the scientific field defines it as an integral process, which starts from the data extraction to the project implementation and monitoring within an organization.

Besides, we can affirm that the classic ML algorithms for DM continue to be applied; however, there are some techniques and algorithms that stand out at present, such as the NN and its extension, the DL, with its algorithms: CNN, GAN and RNN.

Some techniques and algorithms were also identified that, although having origin in past decades, are in the last years when they are of greater application and in a progressive way they are entering in the scientific scope, such as the genetic and evolutionary algorithms, the analysis of feelings, the spatial data mining, the detection of outliers and the recognition and grouping of images.

It was also shown that the types of techniques most used today are clustering and classification.

It was also observed that magazines, standards, surveys and web portals recognized in the scientific field, are migrating from the terminology data mining or data science, always accompanied by the ML. Among these are the KDNuggets portal and the Gartner Magic Quadrant.

Finally, making a comparison with studies dating from 2016 regarding the leading software in the field of data science and the ML, it was observed that some of these companies remain as partners so that the common of the current computer platforms have the implementation of emerging techniques and distinguish themselves by their efficiency in response times, advanced analytics, visualization tools and business intelligence.

Besides, we can consider WEKA as an ally for being open source and remains one of the software with a greater presence in the scientific field.

Without doubt, the massive generation of information that takes place today and the evolution of the research and application fields are those that mark the compass in the incursion of new techniques, optimization of the already existing ones and repowering of the classic ML algorithms for DM and, by default, the implementation in the most outstanding software (Tables 8, 9, 10, 11, 12).

Table 10 Capacities according to the CRISP-DM phases in business/data understanding. *Source*: self-made

FASES CRISP-DM	Business/data understanding			
	WEKA	RapidMiner	TIBCO Spotfire Data Science	Alteryx
Type of file	Excel, SAS, Acces, MRI Nifit, Jso	Excel, SAS, Stata, CSV, WEKA, Access, XML, C4.5, Mdb, texto plano	Excel, SAS, Stata, CSV, Access, XML, Mdb, texto plano	Archivos planos, ESRI, Big data
Open source database	MySQL, PostgreSQL			
Commercial database	JDBC Compatible: Oracle, SQL Server			
Big data	Hadoop, Spark	Hive, Radoop (optimizado para Hadoop), Spark	Spotfire, Jaspersoft	Jaspersoft
Not SQL	Cassandra	MongoDB, Cassandrar	ActiveSpace	Yes

Table 11 Capacities according to the CRISP-DM phases in evaluation/deployment of the data. *Source*: self-made

FASES CRISP-DM	Evaluation/Deployment			
	WEKA	RapidMiner	TIBCO Spotfire Data Science	Alteryx
Display	Yes	RapidMiner Server: web access to reports, results and web processes/services	Consumer and Business Author	Alteryx Visualytics: Tableau, ESRI, MapInfo
Activity displays	Not	RapidMiner Server: web access to reports, results and web processes/services	Analyst	Alteryx Designer

Table 12 DM and ML application areas. *Source*: self-made

Area	Application
Business and Industry	Process modeling and optimization, bottleneck identification
Commerce and Marketing	Through association rules, customer analysis, demand forecasting, market segmentation, targeted marketing is possible
Finance and Banking	Using segmentation techniques, identify new patterns of financial fraud
Medicine and Pharmacy	Recognition and classification of patterns in imaging and diagnostic tests, genetics, prognosis of response and effectiveness of medical treatments
Security	Facial recognition, biometric identifications, unauthorized access to networks
Non-numerical information retrieval	Search and identification of image, video, voice and text from multimedia and web databases
Mining, agriculture and fishing	Identification of areas of use for different crops, fishing or mining in databases of satellite images
Environmental Sciences	Identification of operating models of natural and/or artificial ecosystems (e.g., wastewater treatment plants) to improve their observation, management and/or control
Social Sciences	Study of public opinion flows, voting preferences
City planning	Traffic light planning, mobility and temporal space studies, identification of sociodemographic values
E-Learning	The digital pedagogical model MOOCs (massive open online courses), webinars, virtual classes

References

1. Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press, Cambridge
2. Pérez López C, Santín González D (2007) Minería de datos. Técnicas y herramientas: técnicas y herramientas, 808. Editorial Paraninfo
3. Gutiérrez JA, Molina B (2015) Identificación de técnicas de minería de datos para apoyar la toma de decisiones en la solución de problemas empresariales. Revista Ontare 3(2):33–51. https://doi.org/10.21158/23823399.v3.n2.2015.1440

4. Peralta FC (2014) Proceso de conceptualización del entendimiento del negocio para proyectos de explotación de información. Revista Latinoamericana de Ingeniería de Software 2(5):273–306
5. Azevedo AIRL, Santos MF (2008) KDD, SEMMA and CRISP-DM: a parallel overview. IADS-DM
6. Russell S, Norvig P (2010) Intelligence artificielle: Avec plus de 500 exercices. Pearson Education France, Londres
7. Bucheli H, Thompson W (2014) Statistics and machine learning at scale: new technologies apply machine learning to big data. In: Insights From the Analytics 2014 Conference
8. Simoudis E (1996) Reality check for data mining. IEEE Ann Hist Comput 11(05):26–33
9. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) The KDD process for extracting useful knowledge from volumes of data. Commun ACM 39(11):27–34
10. García-Peñalvo FJ, Montoya MSR (2017) Aprendizaje, Innovación y Competitividad: La Sociedad del Aprendizaje. Revista de Educación a Distancia (RED) (52)
11. Han J, Kamber M, Pei J (2012) 13-data mining trends and research frontiers. Data Mining (Third Edition), ed Boston: Morgan Kaufmann, pp 585–631
12. Viera ÁFG (2017) Técnicas de aprendizaje de máquina utilizadas para la minería de texto. Investigación bibliotecológica 31(71):103–126
13. Xu Z, King I, Lyu MRT, Jin R (2010) Discriminative semi-supervised feature selection via manifold regularization. IEEE Trans Neural Netw 21(7):1033–1047
14. Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. J Artif Intell Res 4:237–285
15. Sutton RS, Barto AG (1998) Introduction to reinforcement learning, vol 135. MIT Press, Cambridge 5:21–22
16. Boada MJL, Boada BL, López VD (2005) Algoritmo de aprendizaje por refuerzo continuo para el control de un sistema de suspensión semi-activa. Revista Iberoamericana de Ingeniería Mecánica 9(2):77
17. Zhang L, Liu P, Gulla JA (2019) Dynamic attention-integrated neural network for session-based news recommendation. Mach Learn 108(10):1851–1875
18. Petticrew M, Roberts H (2008) Systematic reviews in the social sciences: a practical guide. Wiley, New York
19. Eiben AE, Smith JE (2003) Introduction to evolutionary computing. Springer, Berlin ((**Vol. 53, p. 18**))
20. Alcalá R, Gacto MJ, Alcalá-Fdez J (2018) Evolutionary data mining and applications: A revision on the most cited papers from the last 10 years (2007–2017). Wiley Interdiscip Rev Data Min Knowl Discov 8(2):e1239
21. Babi C, Rao MV, Rao VV. Mining frequent patterns from big data sets using genetic algorithm
22. Khan K, Rehman SU, Aziz K, Fong S, Sarasvady S (2014) DBSCAN: Past, present and future. In: The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014). IEEE, pp 232–238
23. Sharma A, Gupta RK, Tiwari A (2016) Improved density based spatial clustering of applications of noise clustering algorithm for knowledge discovery in spatial data. Math Probl Eng 2016
24. Li Y, Xu L, Liu T (2016) Unsupervised change detection for remote sensing images based on object-based MRF and stacked autoencoders. In: 2016 International Conference on Orange Technologies (ICOT). IEEE, pp. 64–67
25. Venugopal N (2019) Sample selection based change detection with dilated network learning in remote sensing images. Sens Imaging 20(1):1–22
26. Haque MA, Mineno H (2018) Contextual outlier detection in sensor data using minimum spanning tree based clustering. In: 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2). IEEE, pp 1–4
27. Bezerra CG, Costa BSJ, Guedes LA, Angelov PP (2016) A new evolving clustering algorithm for online data streams. In: 2016 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS). IEEE, pp 162–168
28. Fernández-Gavilanes M, Álvarez-López T, Juncal-Martínez J, Costa-Montenegro E, González-Castaño FJ (2016) Unsupervised method for sentiment analysis in online texts. Expert Syst Appl 58:57–75
29. Deeva G, De Smedt J, De Koninck P, De Weerdt J (2017) Dropout prediction in MOOCs: a comparison between process and sequence mining. In: International Conference on Business Process Management. Springer, Cham, pp 243–255

30. Bengio Y, LeCun Y (2007) Scaling learning algorithms towards AI. Large-scale Kernel Mach 34(5):1–41
31. Wang K, Gou C, Duan Y, Lin Y, Zheng X, Wang FY (2017) Generative adversarial networks: introduction and outlook. IEEE/CAA J Autom Sin 4(4):588–598
32. Ciresan DC, Meier U, Masci J, Gambardella LM, Schmidhuber J (2011) Flexible, high performance convolutional neural networks for image classification. In: Twenty-Second International Joint Conference on Artificial Intelligence
33. Yao G, Lei T, Zhong J (2019) A review of convolutional-neural-network-based action recognition. Pattern Recognit Lett 118:14–22
34. Sharma N, Jain V, Mishra A (2018) An analysis of convolutional neural networks for image classification. Proc Comput Sci 132:377–384
35. Zou L, Yu S, Meng T, Zhang Z, Liang X, Xie Y (2019) A technical review of convolutional neural network-based mammographic breast cancer diagnosis. Comput Math Methods Med
36. Lipton ZC, Berkowitz J, Elkan C (2015) A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019
37. Gauthier J (2014) Conditional generative adversarial nets for convolutional face generation. In: Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester 2014(5), 2
38. Patel J, Pandya MS, Shah V (2018) Review on generative adversarial networks. Tech Innov Modern Eng Sci 7(4):2018
39. Hidasi B, Karatzoglou A, Baltrunas L, Tikk D (2015) Session-based recommendations with recurrent neural networks. arXiv:1511.06939
40. Tan YK, Xu X, Liu Y (2016) Improved recurrent neural networks for session-based recommendations. In: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, pp 17–22
41. Blei DM, Lafferty JD (2006) Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning, pp 113–120
42. Chen SA, Tangkaratt V, Lin HT, Sugiyama M (2019) Active deep Q-learning with demonstration. Mach Learn 1–27
43. TIBCO, TIBCO (2017) Product Documentation, 74, 84, TIBCO
44. KDNuggets, Kdnuggets (2019) https://kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html
45. Bermúdez JAG, Ramirez Á MA (2010) Análisis para predicción de ventas utilizando minería de datos en almacenes de ventas de grandes superficies (Doctoral dissertation, Universidad Tecnológica de Pereira. Facultad de Ingenierías Eléctrica, Electrónica, Física y Ciencias de la Computación. Ingeniería de Sistemas y Computación)
46. RapidMiner, RapidMiner Studio Manual (2014) https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf
47. WEKA (2019) Weka 3: Machine Learning Software in Java, WEKA, 2019, http://www.cs.waikato.ac.nz/ml/weka/
48. González FJG, Aguilera SG, Jurado JAM (2013) Aplicación de técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA). Granada. Obtenido de https://www.ugr.es/doctoeio/TFM2013/TFM/_GarciaGonzalezFrancisco.pdf
49. Rbigui H, Cho C (2017) The state-of-the-art of business process mining challenges. Int J Bus Process Integr Manag 8(4):285–303
50. Garcia AM, Charte F, González P, Carmona CJ, del Jesus MJ (2016) Subgroup discovery with evolutionary fuzzy systems in R: the SDEFSR package. R J 8(2):307
51. Deshpande S, Doke M, Deshpande A, Chaudhari AN (2017) Expert system for retrieval of documents using evolutionary approaches incorporating clustering. In: 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), vol 2. IEEE, pp 414–418
52. Shi Y, Du S (2017) Manifold regularized robust unsupervised feature selection for image clustering. In: 2017 36th Chinese Control Conference (CCC). IEEE, pp 11161–11165
53. Sarala R, Saravanan V. Spatio-temporal pattern discovery using machine learning random forests approach
54. Suo Y, Liu T, Jia X, Yu F (2018) Application of clustering analysis in brain gene data based on deep learning. IEEE Access 7:2947–2956

55. Sumalatha V, Santhi R (2018) An improved Bayes classification approach to reduce affliction of Juvenile. In: 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE, pp 1–4
56. Femi PS, Vaidyanathan SG (2018) Comparative study of outlier detection approaches. In: 2018 International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, pp 366–371
57. Li Y (2018) Feature extraction and learning effect analysis for MOOCs users based on data mining. Int J Emerg Technol Learn iJET 13(10):108–120
58. Abd El-Jawad MH, Hodhod R, Omar YM (2018) Sentiment analysis of social media networks using machine learning. In: 2018 14th International Computer Engineering Conference (ICENCO). IEEE, pp 174–176

Authors and Affiliations

Gilda Taranto-Vera[1] · Purificación Galindo-Villardón[1] · Javier Merchán-Sánchez-Jara[1] · Julio Salazar-Pozo[1] · Alex Moreno-Salazar[1,2] · Vanessa Salazar-Villalva[1,2]

Purificación Galindo-Villardón
pgalindo@usal.es

Javier Merchán-Sánchez-Jara
javiermerchan@usal.es

Julio Salazar-Pozo
julio_salazar@usal.es

Alex Moreno-Salazar
amorenos@usal.es

Vanessa Salazar-Villalva
vanessa.salazar@usal.es

[1] Universidad de Salamanca, Salamanca, Spain

[2] Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador