Machine Learning. Aprendizaje Automático.

Métodos, herramientas

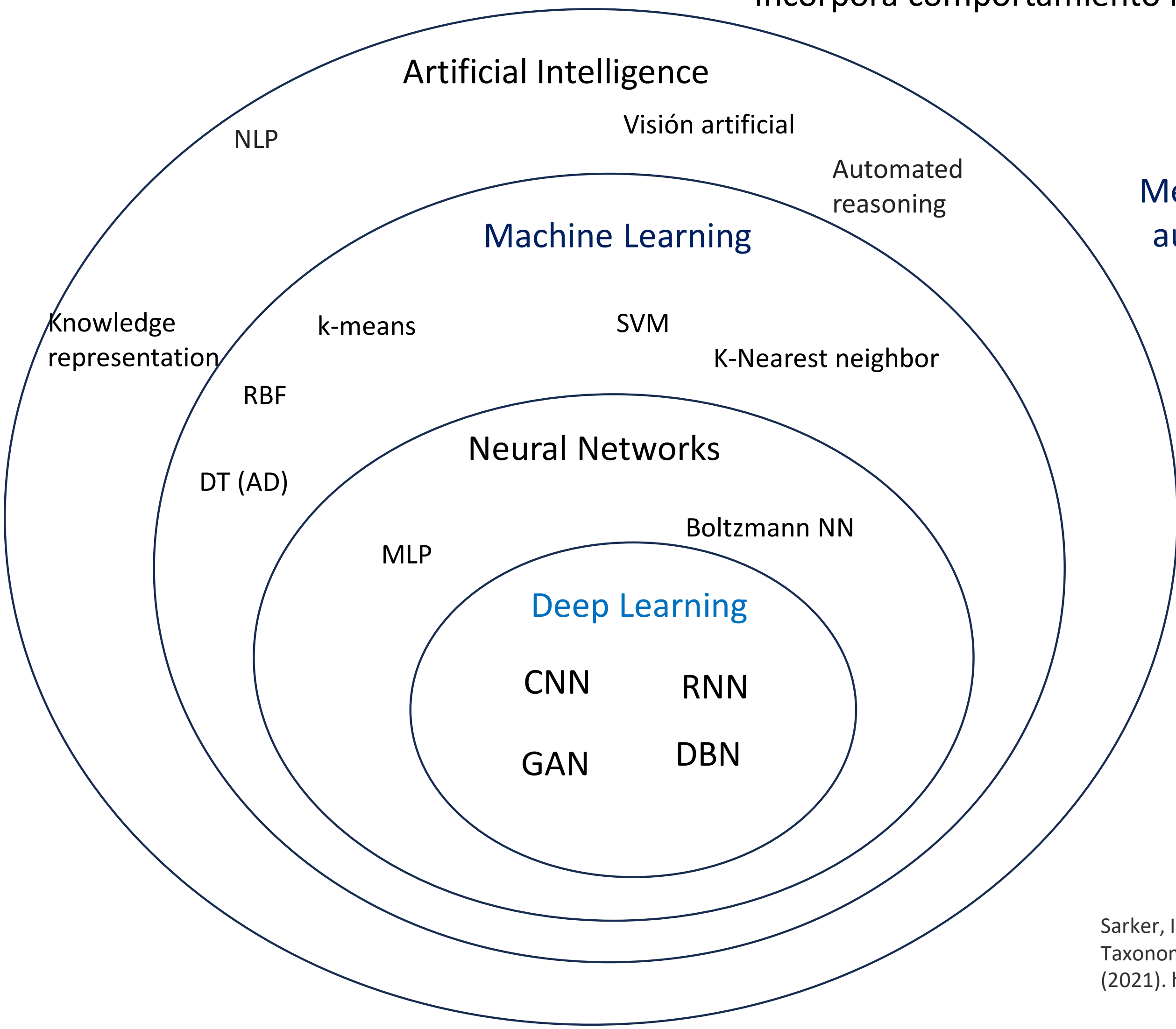
Prof. Dra. Sonia I. Mariño
simarinio@yahoo.com
2025

Indice

Introducir en

- Inteligencia Artificial (IA), Ciencias de Datos (CD), Machine Learning (ML)
- Aprendizaje en CD, ML.
- Algunos métodos, metodologías, pieplines, workflow aplicables en CD, ML
- Algunos recursos

Incorpora comportamiento humano e inteligencia a máquinas o sistemas



Métodos que aprenden de datos o experiencia, automatización en la construcción de modelos

- Multilayer Perceptrons (MLPs)
- Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs)
- Long Short Term Memory Networks (LSTMs)
- Generative Adversarial Networks (GANs)
- Radial Basis Function Networks (RBFNs)
- Self Organizing Maps (SOMs)
- Deep Belief Networks (DBNs)
- Restricted Boltzmann Machines(RBMs)
- Autoencoders
- LLM (Large Language Model)
- Otros

Computación utilizando RN multi-layer

Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN COMPUT. SCI.* **2**, 420 (2021). https://doi.org/10.1007/s42979-021-00815-1

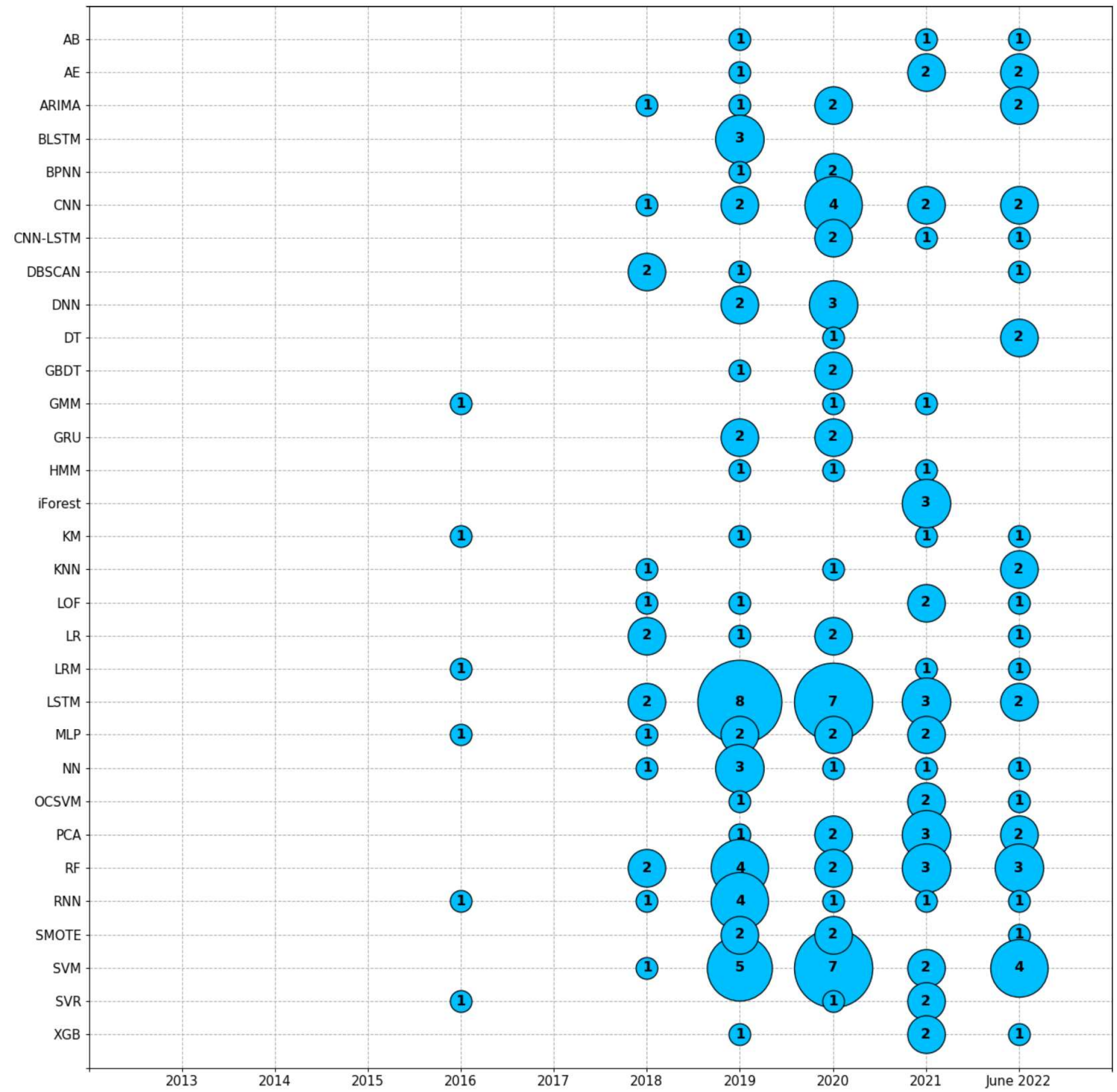
Ejemplo de uso de una tecnología de IA

Figure 5 shows how many times a data science method appeared over the years of publication. Long short-term memory (LSTM) networks were the method that most appeared in the corpus, with 22 occurrences. Then, support vector machine (SVM) had 19 occurrences. Next, the random forest (RF) method appeared 14 times. The years 2019, 2020, and 2021 presented the highest concentration of data science methods.

Fuente datos
ACM, IEEE, Scopus, Springer, and Wiley

Arruda, H.M.; Bavaresco, R.S.; Kunst, R.; Bugs, E.F.; Pesenti, G.C.; Barbosa, J.L.V. Data Science Methods and Tools for Industry 4.0: A Systematic Literature Review and Taxonomy. Sensors 2023, 23, 5010.

https://doi.org/10.3390/s23115010



Conocimiento y aprendizaje en ML

Aprendizaje

responde a diversos fenómenos.

- Perfeccionar una habilidad.
- Adquirir conocimiento.

Modalidades (algunos autores proponen 3 clases de aprendizaje y otros proponen 4 clases de aprendizaje)

- Aprendizaje supervisado
- Aprendizaje no supervisado
- Aprendizaje semi-supervisado
- Aprendizaje por refuerzo

Aprendizaje Supervisado

Algunos algoritmos más utilizado en ML

- Modelos lineales para regresión.
- Modelos lineales para clasificación.
- Árboles y bosques. Kernels y máquinas de soporte vectorial.
- Redes neuronales. Backpropagation.
- Redes bayesianas.

- Se disponen de datos “etiquetados” y sus correspondientes valores de salida
- El algoritmo aprende a través de un entrenamiento con un conjunto de datos históricos / conocido.
- En procesos posteriores puede predecir o clasificar para proponer solución a un problema
- Resuelve problemas de clasificación y regresión

Aprendizaje no Supervisado

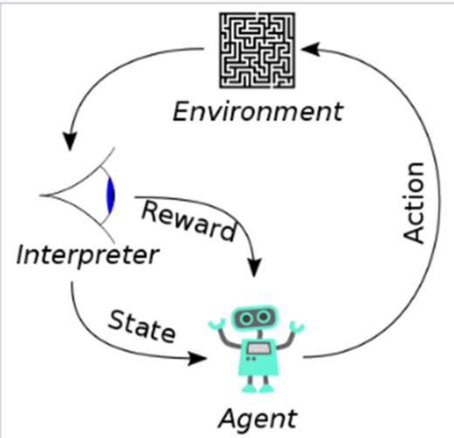
- Clustering.
- K-Means. Mezcla de Gaussianas. HDBscan.
- Análisis en componentes principales (PCA). PCA probabilístico. PCA Bayesiano. PCA con núcleos. Análisis de componentes independientes.
- Markov Chain Monte Carlo. Metropolis Hastings. Muestreo con ensambles. Inferencia aproximada. Naive Bayes. Métodos variacionales Bayesianos.

- Se carece de datos "etiquetados" para el entrenamiento y se desconoce los datos de salida correspondientes a cada instancia o input.
- Algoritmos exploratorios, descubrir patrones o estructuras en los datos.
- Finalidad: agrupar los datos con similitudes
- Problemas no-supervisados, presentan alta dimensionalidad dado que se caracteriza el espacio de las entradas,
  - Se requiere aplicar técnicas para reducir la dimensionalidad.

Aprendizaje semi-supervisado

- Combina algoritmos de Aprendizaje Supervisado para etiquetar puntos de datos con etiquetas conocidas, y algoritmos de Aprendizaje no Supervisado para agrupar puntos de datos.
- Se aplica a problemas costosos en tiempo o procesamiento.
- Ejemplos.
  - Algoritmo: Deep Belief Networks (DBN) – o Redes de Creencia Profunda-, compuestas de redes simples denominadas Restricted Boltzmann Machines (RBS) [entrenamiento no supervisado de manera secuencial], y se continua con entrenamiento supervisado.
  - Etiquetar algunas personas en fotos y aplicar procesos de reconocimiento *a posteriori*

Aprendizaje por refuerzo



The typical framing of a Reinforcement Learning (RL) scenario: an agent takes actions in an environment, which is interpreted into a reward and a representation of the state, which are fed back into the agent.

Aprendizaje que mejora la respuesta del modelo usando un proceso de retroalimentación.

Aprendizaje a partir de la observación del mundo en que interviene.

Información de entrada, es el feedback o retroalimentación del mundo exterior como respuesta a sus acciones.

El aprendizaje se basa en ensayo-error.

Puede ser entendido como AS

Algoritmos: Q-learning, Deep Q-learning

Aplicaciones: entretenimiento, salud, otros

ML. Aprendizaje y aplicaciones

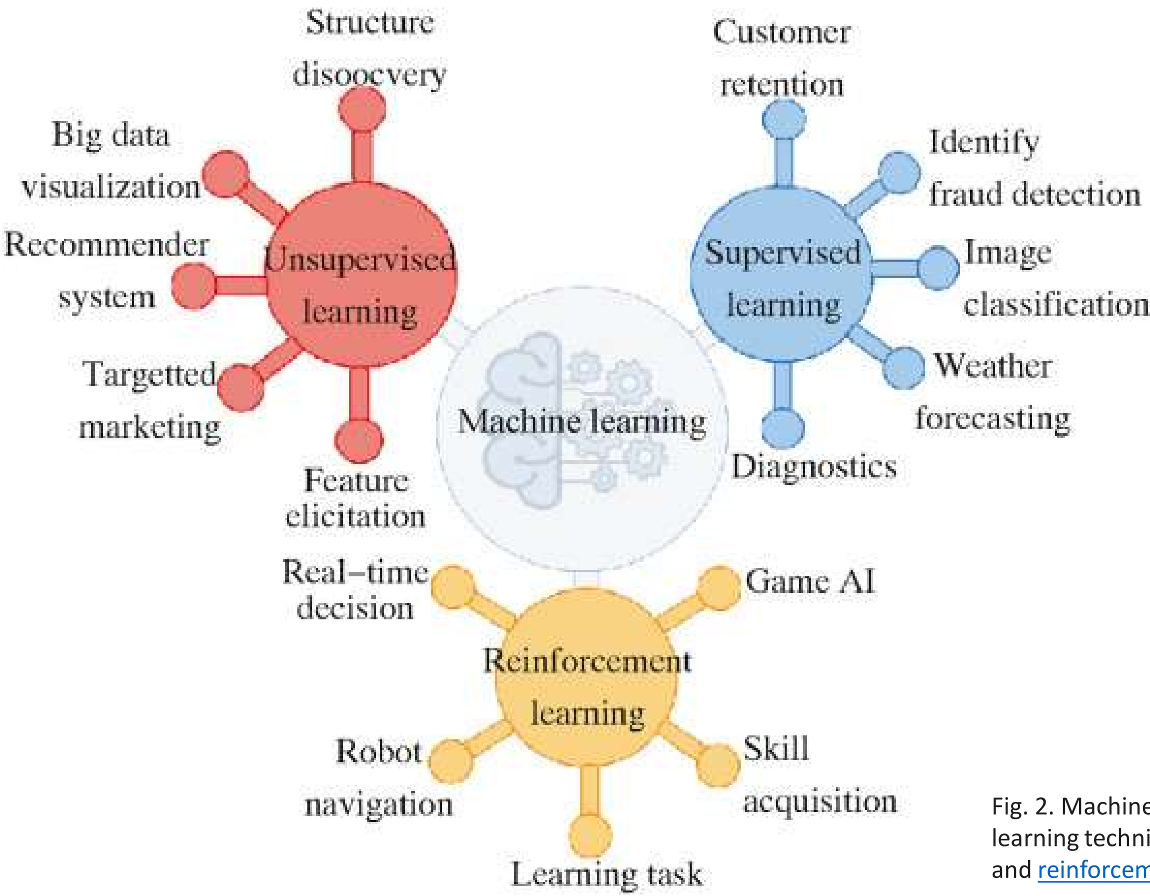


Fig. 2. Machine learning algorithms classification. Classification of the main machine learning techniques, namely supervised learning, unsupervised learning, and reinforcement learning with some examples.

https://www.sciencedirect.com/science/article/pii/S2666764921000485

TABLE 2.
DNN network comparison table.

Network Type	Architecture	Network Model	Training Type	Training Algorithm	Implementation Sample	Common Application	Popular Dataset Sample	DL Framework (sample)
Feedforward Neural Network	CNN	Discriminative	Supervised	Gradient Descent based Backpropagation	Siamese Network, Deep CNN	Image recognition/classification	MNIST	TensorFlow, Caffe, Theano, Torch, Deeplearning4j, Microsoft Cognitive Toolkit, Keras, MXNet, PyTorch
	Residual Network	Discriminative	Supervised	Gradient Descent based Backpropagation	Deep ResNet; HighwayNet; DenseNet	Image recognition	ImageNet	TensorFlow, PyTorch, Keras
	Autoencoder	Generative	Unsupervised	Backpropagation	Sparse Autoencoders, Variational Autoencoders	Dimensionality Reduction; Encoding	MNIST	TensorFlow, Deeplearning4j, Keras
	Adversarial Networks	Generative & Discriminative	Unsupervised	Backpropagation	Generative Adversarial Network	Generate realistic fake data; Reconstruction of 3D models; Image improvement	CIFAR10	TensorFlow, Keras
	RBM	Generative with Discriminative finetuning	Unsupervised	Gradient Descent based Contrastive divergence	Deep Belief Network; Deep Boltzmann Machine	Dimensionality Reduction; Feature learning; Topic modeling	MNIST	TensorFlow, Deeplearning4j, Keras, MXNet, Theano, Torch
Recurrent Neural Network	LSTM	Discriminative	Supervised	Gradient Descent & Backpropagation through Time	Deep RNN, Gated Recurrent Unit (GRU), Neural Machine Translation (NMT)	Natural Language Processing; Language Translation	MNIST Stroke Sequence	TensorFlow, Caffe, Theano, Torch, Deeplearning4j, Microsoft Cognitive Toolkit, Keras, MXNet, PyTorch
Radial Basis Function NN	RBF Network	Discriminative	Supervised and Unsupervised	K-means Clustering; Least Square Function	Radial Basis Function NN	Function approximation; Time series prediction	Fisher's Iris data set	TensorFlow
Kohonen Self Organizing NN	Nodes arranged in hexagonal or rectangular grid	Generative	Unsupervised	Competitive Learning	Kohonen Self Organizing NN	Dimensionality Reduction; Optimization problems; Clustering analysis	SPAMbase	TensorFlow

IEEE Xplore Full-Text PDF:

Metodologías

Algunas metodologías / métodos:

- KDD
- *CRISP-DM*
- *SEMMA*
- *Workflows, Pipelines…*

Metodologías de minería de datos (MD)

KDD process (Knowledge Discovery in Databases=.

- Desarrollada en 1989, por Gregory Piatetsky-Shapiro
- Descubre patrones al procesar datos sin procesar, analizar la información en busca de datos necesarios e interpretar los resultados.

DATOS

Selección

Entender el dominio de aplicación, el problema a resolver, y cuales son los objetivos.

Datos (muestras)

Preprocesamiento

Recolección y limpieza de muestra, atrbutos faltantes y atipico

Datos Procesados

Transformación

Reducción de datos y explotacion de formas de representación y normalización

Datos Transformados

Extracción de Datos

Busqueda de patrones y reglas

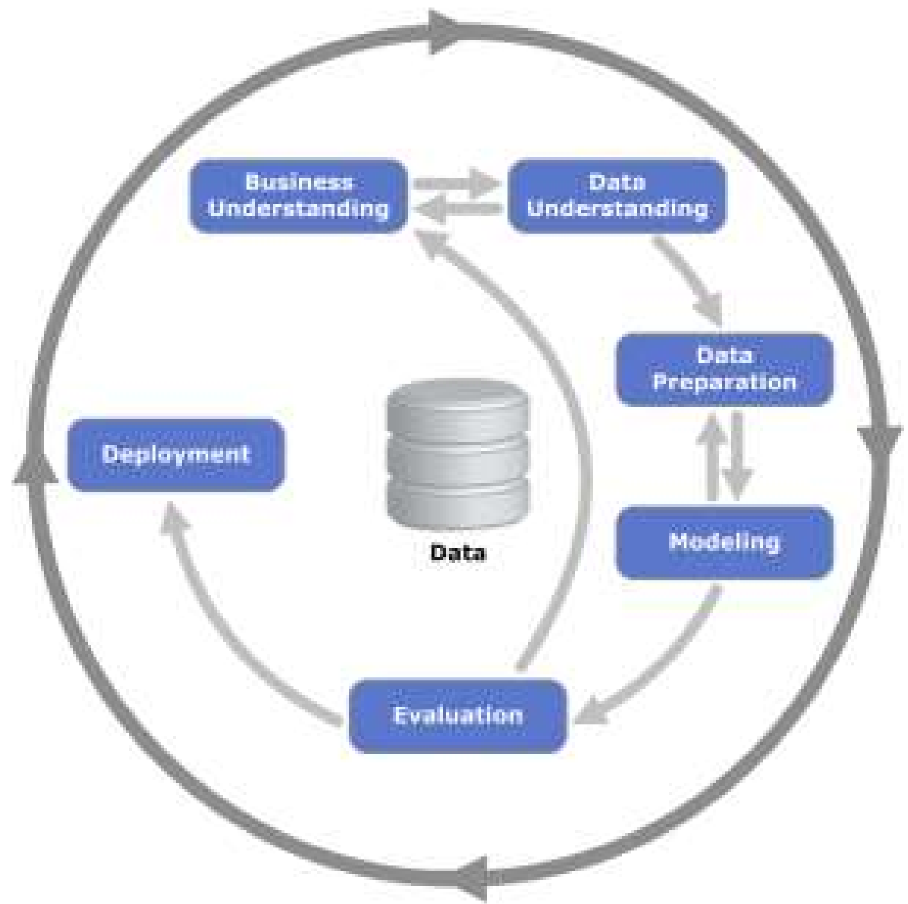
Patrones Información

Interpretación y validación

CONOCIMIENTO

CRISP-DM - Cross Industry Standard Process for Data Mining

- Perspectiva de **metodología**, descripciones de las fases de un proyecto, las tareas en cada fase y explica las relaciones entre las tareas.
- Perspectiva de **modelo de proceso**, resumen del ciclo de minería de datos.



CRISP-DM, http://www.crisp-dm.org/CRISPWP-0800.pdf

CRISP-DM - Cross Industry Standard Process for Data Mining

1. Business Understanding
    - Determine Business Objectives
    - Assess Situation
    - Determine Data Science Goals
    - Produce Project Plan
2. Data Understanding
    - Collect Initial Data
    - Describe Data
    - Explore Data
    - Verify Data Quality
3. Data Preparation
    - Select Data
    - Clean Data
    - Construct Data
    - Integrate Data
    - Format Data
4. Modeling
    - Select Modeling Technique
    - Generate Test Design
    - Build Model
    - Assess Model
5. Evaluation
    - Evaluate Results
    - Review Process
    - Determine Next Steps
6. Deployment
    - Plan Deployment
    - Plan Monitoring & Maintenance
    - Produce Final Report
    - Review Project

Abstract

CRISP-DM is the de-facto standard and an industry-independent process model for applying data mining projects. Twenty years after its release in 2000, we would like to provide a systematic literature review of recent studies published in IEEE, ScienceDirect and ACM about data mining use cases applying CRISP-DM. We give an overview of the research focus, current methodologies, best practices and possible gaps in conducting the six phases of CRISP-DM. The main findings are that CRISP-DM is still a de-factor standard in data mining, but there are challenges since the most studies do not foresee a deployment phase. The contribution of our paper is to identify best practices and process phases in which data mining analysts can be better supported. Further contribution is a template for structuring and releasing CRISP-DM studies.

Peer-review under responsibility of the scientific committee of the CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2020

Keywords: CRISP-DM; Literature Review; Data Mining; Process Methodology; Deployment

Table 1: CRISP-DM process model descriptions [10].

Phase	Short description
Business Understanding	The business situation should be assessed to get an overview of the available and required resources. The determination of the data mining goal is one of the most important aspect in this phase. First the data mining type should be explained (e. g. classification) and the data mining success criteria (like precision). A compulsory project plan should be created.
Data understanding	Collecting data from data sources, exploring and describing it and checking the data quality are essential tasks in this phase. To make it more concrete, the user guide describe the data description task with using statistical analysis and determining attributes and their collations.
Data preparation	Data selection should be conducted by defining inclusion and exclusion criteria. Bad data quality can be handled by cleaning data. Dependent on the used model (defined in the first phase) derived attributes have to be constructed. For all these steps different methods are possible and are model dependent.
Modeling	The data modelling phase consists of selecting the modeling technique, building the test case and the model. All data mining techniques can be used. In general, the choice is depending on the business problem and the data. More important is, how to explain the choice. For building the model, specific parameters have to be set. For assessing the model it is appropriate to evaluate the model against evaluation criteria and select the best ones.
Evaluation	In the evaluation phase the results are checked against the defined business objectives. Therefore, the results have to be interpreted and further actions have to be defined. Another point is, that the process should be reviewed in general.
Deployment	The deployment phase is described generally in the user guide. It could be a final report or a software component. The user guide describes that the deployment phase consists of planning the deployment, monitoring and maintenance.

CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories

Fernando Martínez-Plumed, Lidia Contreras-Ochando, Cèsar Ferri, José Hernández-Orallo, Meelis Kull, Nicolas Lachiche, María José Ramírez-Quintana and Peter Flach

Abstract—CRISP-DM (CRoss-Industry Standard Process for Data Mining) has its origins in the second half of the nineties and is thus about two decades old. According to many surveys and user polls it is still the *de facto* standard for developing data mining and knowledge discovery projects. However, undoubtedly the field has moved on considerably in twenty years, with *data science* now the leading term being favoured over *data mining*. In this paper we investigate whether, and in what contexts, CRISP-DM is still fit for purpose for data science projects. We argue that if the project is goal-directed and process-driven the process model view still largely holds. On the other hand, when data science projects become more exploratory the paths that the project can take become more varied, and a more flexible model is called for. We suggest what the outlines of such a trajectory-based model might look like and how it can be used to categorise data science projects (goal-directed, exploratory or data management). We examine seven real-life exemplars where exploratory activities play an important role and compare them against 51 use cases extracted from the NIST Big Data Public Working Group. We anticipate this categorisation can help project planning in terms of time and cost characteristics.

Index Terms—Data Science Trajectories, Data Mining, Knowledge Discovery Process, Data-driven Methodologies.

TKDE_Data_Science_Trajectories_PF.pdf (bris.ac.uk)

SEMMA

SEMMA

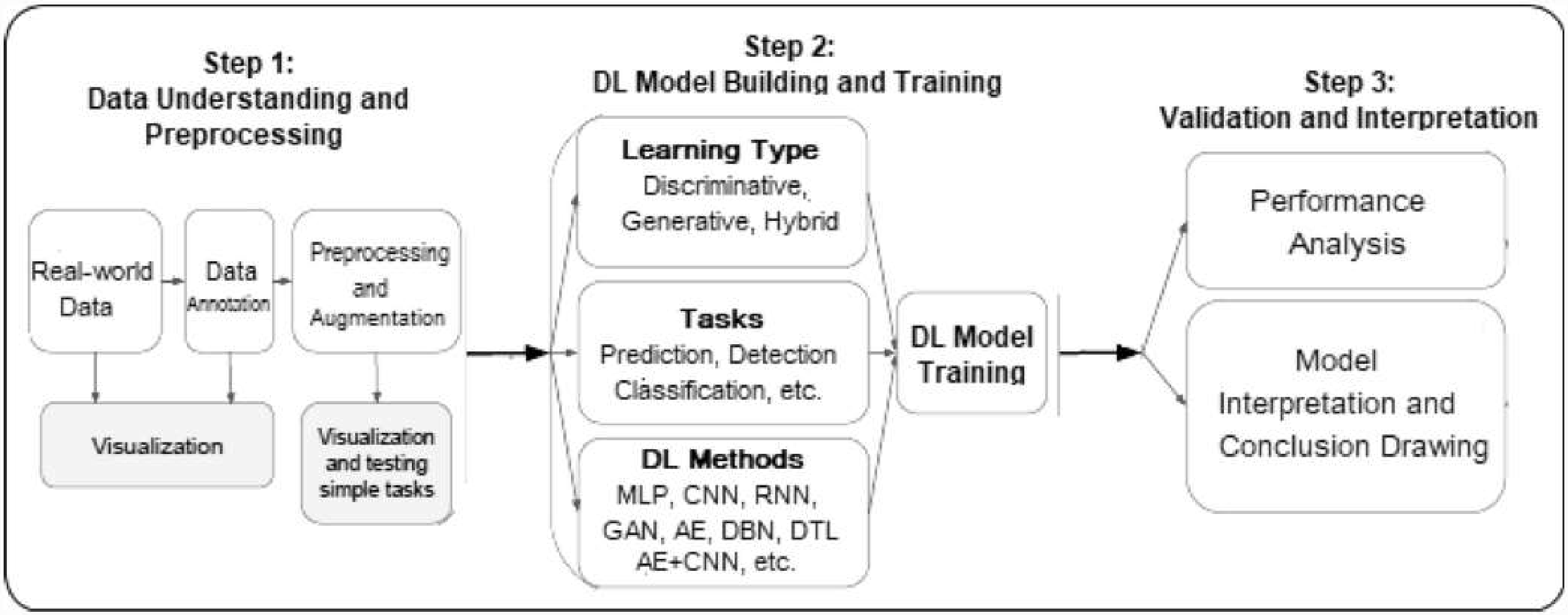
- Sample, Explore, Modify, Model, and Assess.
- Lista de pasos secuenciales desarrollados por SAS Instituto.
- Guía la implementación de aplicaciones de minería de datos.

Sample' (muestreo)	Explore' (exploración)	Modify' (modifica-ción):	Model' (modelado	Assess (evaluación)
captura de datos, y eventualmente creación de tablas u otras estructuras que los contengan, Considerar el tamaño de las muestras para disponer de un conjunto de datos representativo y suficientemente para el procesamiento.	entender los datos disponibles, se aplica a través de la visualización, el agrupamiento, etc para observar las relaciones, tendencias y otra información proporcione conocimiento sobre los datos y el fenómeno subyacente.	trabajo sobre los datos para su uso en el posterior modelado. Se transforman y seleccionan o crean variables a partir de los datos , otras	determina el modelo más adecuado (el que mejor predice la o las variables de salida a partir de la o las variables de entrada), seleccionando entre las familias disponibles (redes neuronales, árboles de decisión, regresión logística, etc) y afinando el modelo en la o las opciones seleccionadas	evaluar el funcionamiento del modelo en su conjunto en cuanto a fiabilidad, utilidad etc

SEMMA, http://www.sas.com/offices/europe/uk/technologies/analytics/datamini ng/miner/semma.html

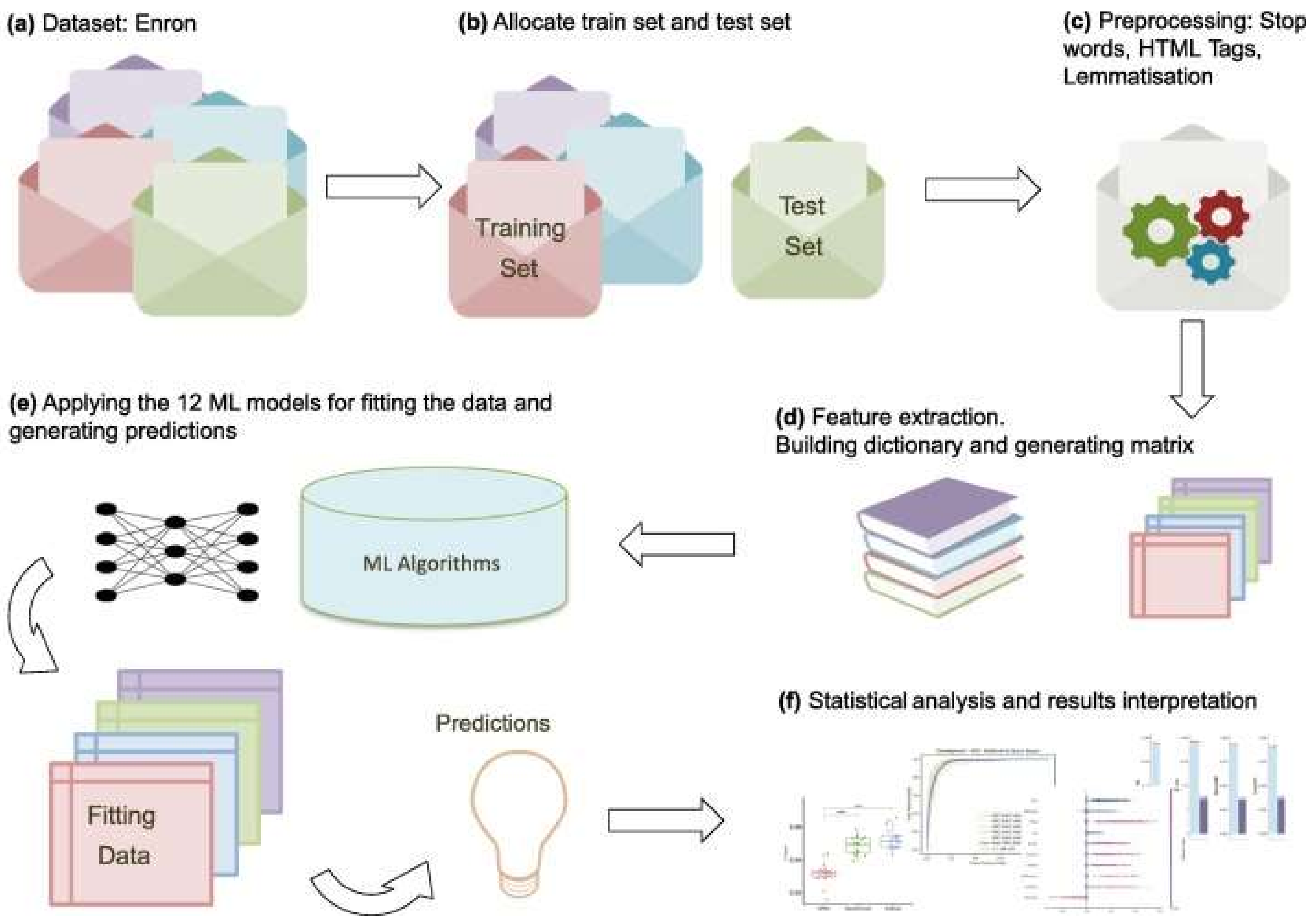
Fig. 4

From: Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions



A typical DL workflow to solve real-world problems, which consists of three sequential stages (i) data understanding and preprocessing (ii) DL model building and training (iii) validation and interpretation

Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN COMPUT. SCI.* **2**, 420 (2021). https://doi.org/10.1007/s42979-021-00815-1



We proposed and tested a pipeline to compare and explain the classification outcomes of 12 machine learning models. We applied the pipeline for optimising and testing the models in a spam filtering context, with lemmatisation and noise-reduction techniques as preprocessing steps. The pipeline, which we make publicly available, was developed to compare the performance of the classifiers in terms of precision, recall, F-score, and ROC curves.

Annalisa Occhipinti, Louis Rogers, Claudio Angione, A pipeline and comparative study of 12 machine learning models for text classification, Expert Systems with Applications, vol 201, 2022, https://doi.org/10.1016/j.eswa.2022.117193.22005802)

ML pipeline

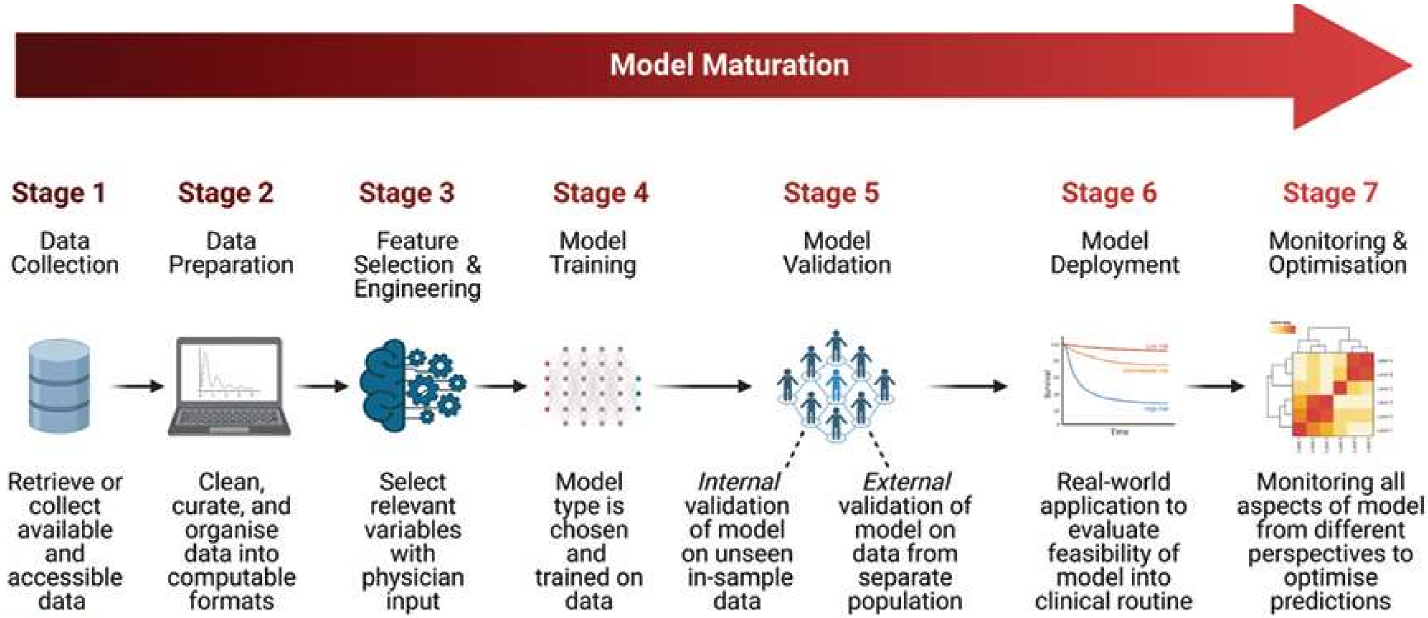


Figure 1 Machine learning pipeline.

The key stages of the ML pipeline that models must traverse, from initial in- silico (computer- based) development to real- world deployment, comprise the following6 (figure 1): (1) data collection; (2) data preparation; (3) feature selection and engineering; (4) model training; (5) model validation, both internal and external; (6) deployment of the model within a working application; and (7) post- deployment monitoring and optimisation of the application. During the development phase (stages 1–3), researchers collect, clean and transform data into computable formats and select relevant features as model inputs. The model is then iteratively improved through several training cycles against static, retrospective datasets (stage 4). In stage 5, the model undergoes two processes of validation: internal validation for accuracy and reproducibility against a random sample from the original training dataset ('hold out' sample); and external validation, whereby researchers validate the model on a new external dataset set derived from previously unencountered patients using the same performance metrics. In stage 6, the model is subject to prospective validation using live (or near- live) dynamic data in a form reflecting its future real- world deployment, integrated into a prototype application, and evaluated for its feasibility in clinical workflows. Then, it is assessed for its clinical utility within clinical trials, which compares application- guided patient care and outcomes with the current standard of care. Finally, stage 7 entails monitoring the effectiveness and safety of the model over its life cycle using surveillance data.

Navigating the machine learning pipeline: a scoping review of inpatient delirium prediction models

ML pipeline

Both pipelines follow a common structure (see Fig. 2): (i) The ingestion step loads the specified raw source data. (ii) The split step splits the ingested dataset into a training dataset for model training, a validation dataset for model performance evaluation and tuning, and a test dataset for model performance evaluation. (iii) The transformation step uses the training dataset to fit a transformer that performs the defined transformations. The transformer is then applied to the training dataset and the validation dataset, creating transformed datasets that are used by subsequent steps for estimator training and model performance evaluation. (iv) The training step uses the transformed training dataset to fit an AutoML or user-defined estimator. The estimator is then joined with the fitted transformer to create a model. Finally, this model is evaluated against the transformed training and validation datasets to compute performance metrics. (v) The evaluation step evaluates the trained model on the test dataset, computing performance metrics and model explanations. Resulting performance metrics are compared against defined thresholds indicating whether a subsequent iteration through the transformation and training steps is necessary.

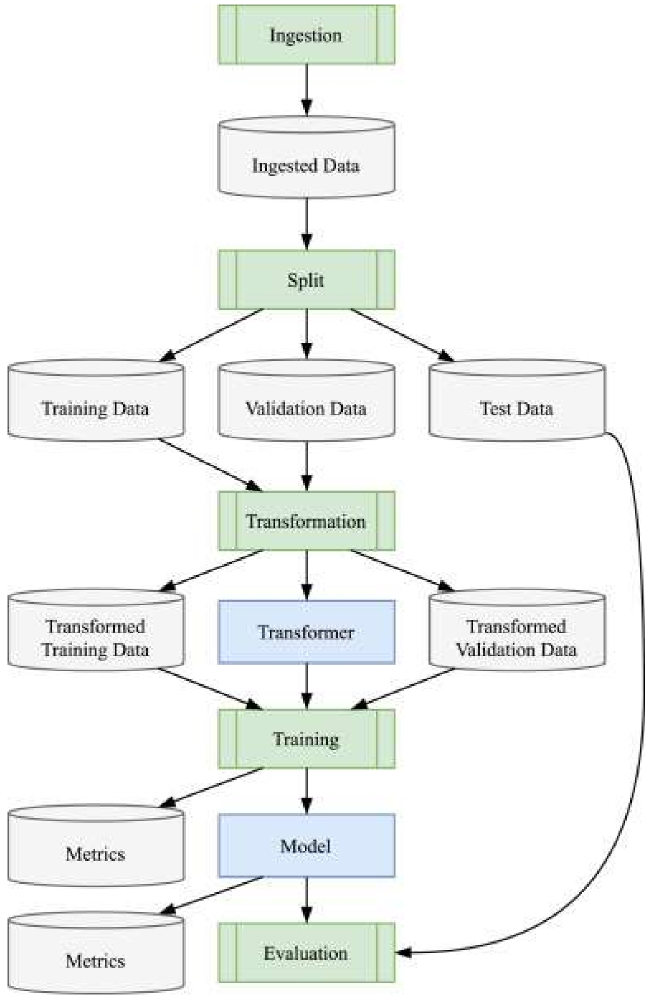
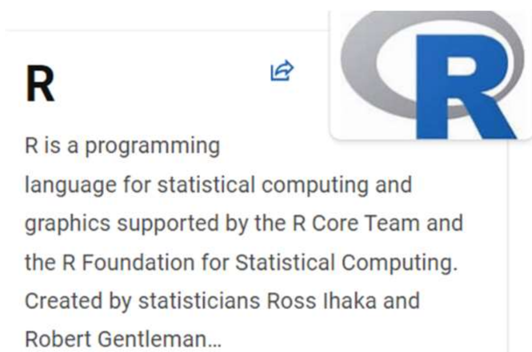


Fig. 2. Pipeline structure of the example classification and regression pipelines.

[Capturing end-to-end provenance for machine learning pipelines - ScienceDirect]

Recursos



https://www.python.org/psf-landing/

Algunos Recursos

Figure 6 shows the software tools grouped by years. The Python programming language was the most used tool, appearing in 20 papers, followed by Keras, which appeared in 15 papers, and Tensorflow which appeared in 13 articles.

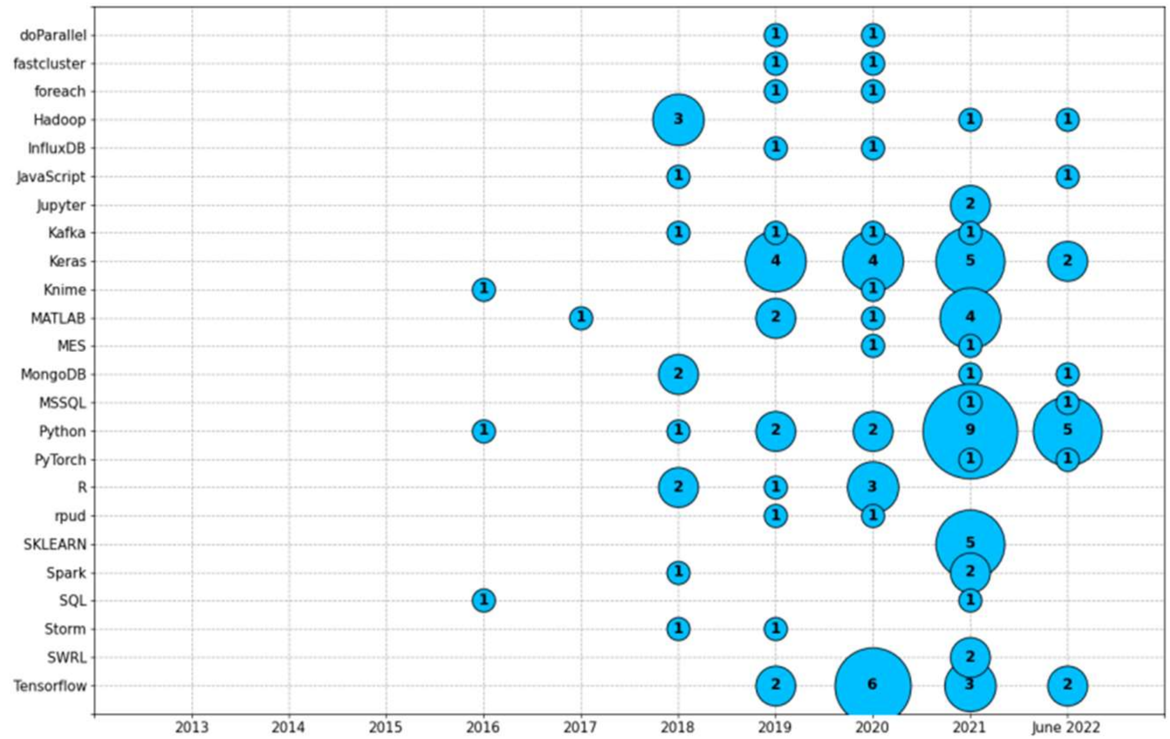


Figure 6. Software tools grouped by year. The definition of each tool is in Table A3. Python was the tool with the most occurrences (20), followed by Keras (15), and Tensorflow (13). For a better visualization, only tools with more than one occurrence appear in the picture.

Citation: Arruda, H.M.; Bavaresco, R.S.; Kunst, R.; Bugs, E.F.; Pesenti, G.C.; Barbosa, J.L.V. Data Science Methods and Tools for Industry 4.0: A Systematic Literature Review and Taxonomy. Sensors 2023, 23, 5010.

https://doi.org/10.3390/s23115010

Fuente datos
ACM, IEEE, Scopus, Springer, and Wiley

TABLE 1. **Popular deep learning frameworks and libraries.**

Framework	Institution	License	1st Release
Caffe	Berkeley AI Research	BSD / Free	2015
Microsoft Cognitive Toolkit	Microsoft	MIT License / Free	2016
Gluon	AWS and Microsoft	Open Source	2017
Keras	Individual Author	MIT License / Free	2015
MXNet	Apache Software Foundation	Apache 2.0 / Free	2015
TensorFlow	Google Brain	Apache 2.0 / Free	2015
Theano	University of Montreal	BSD / Free	2008
Torch	Ronan Collobert et al.	BSD / Free	2002
PyTorch	Facebook	BSD / Free	2016
Chainer	Preferred Networks	BSD / Free	2015
Deeplearning4j	Adam Gibson et al.	Apache 2.0 / Free	2014

Licencia Pública General de GNU (GPL) es una licencia de software libre que se utiliza para proteger la libertad de los usuarios finales de software.

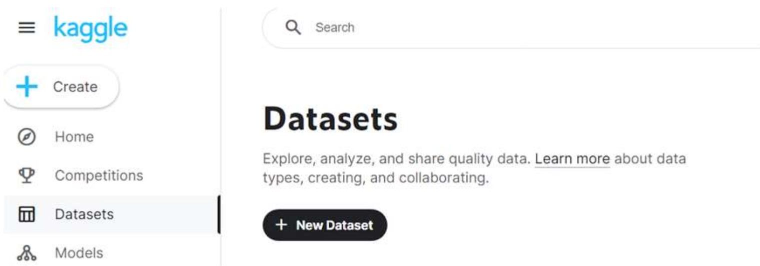
BSD: licencia Berkeley

Apache / MIT
Licencias permisivas NO o exige que las obras derivadas (versiones modificadas) del software se distribuyan usando la misma licencia

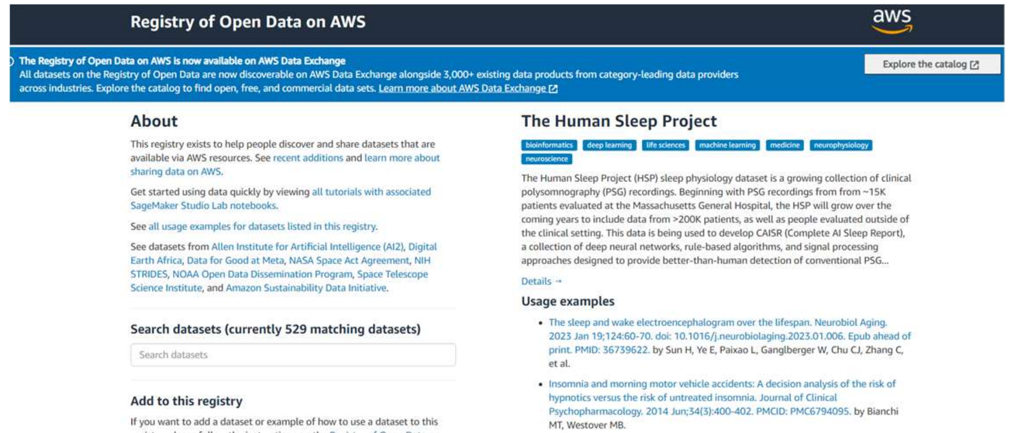
Open Source, otorga a usuario acuerdo legal que define los términos y condiciones bajo los cuales el software puede ser utilizado, modificado y distribuido.

10.1109/ACCESS.2019.2912200

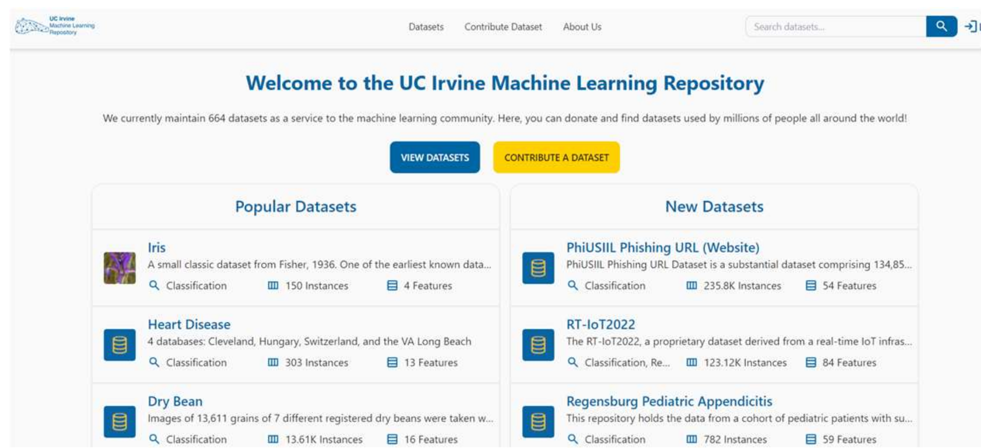
Recursos. Datasets



[Find Open Datasets and Machine Learning Projects | Kaggle]



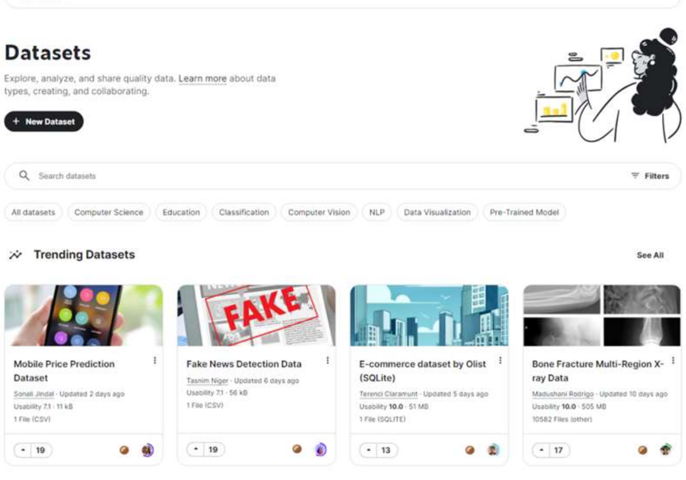
[Registry of Open Data on AWS]



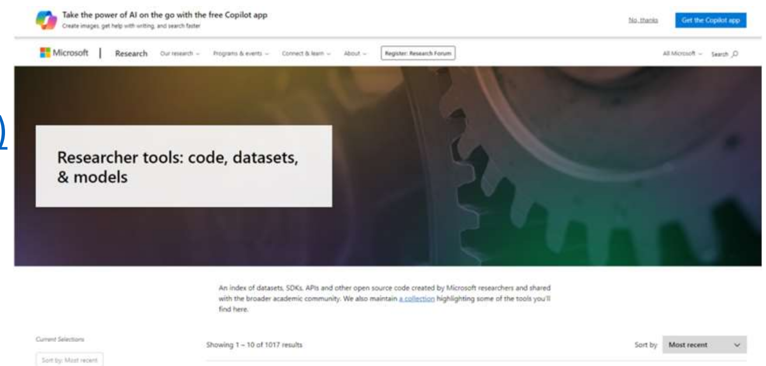
[Home - UCI Machine Learning Repository]

Recursos. Datasets

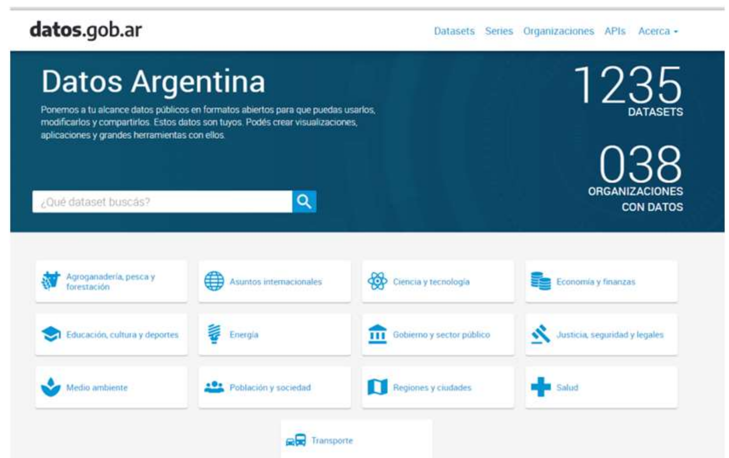
[Portal de Datos Abiertos (ciudaddecorrientes.gov.ar)](ciudaddecorrientes.gov.ar)



https://www.kaggle.com/datasets



[Researcher tools: code, datasets, & models - Microsoft Research](#)



[Datos Argentina](#)

https://catalogo.datos.gba.gob.ar/dataset