Diplomatura Universitaria en Ciencia de Datos

https://exa.unne.edu.ar/diplomatura/

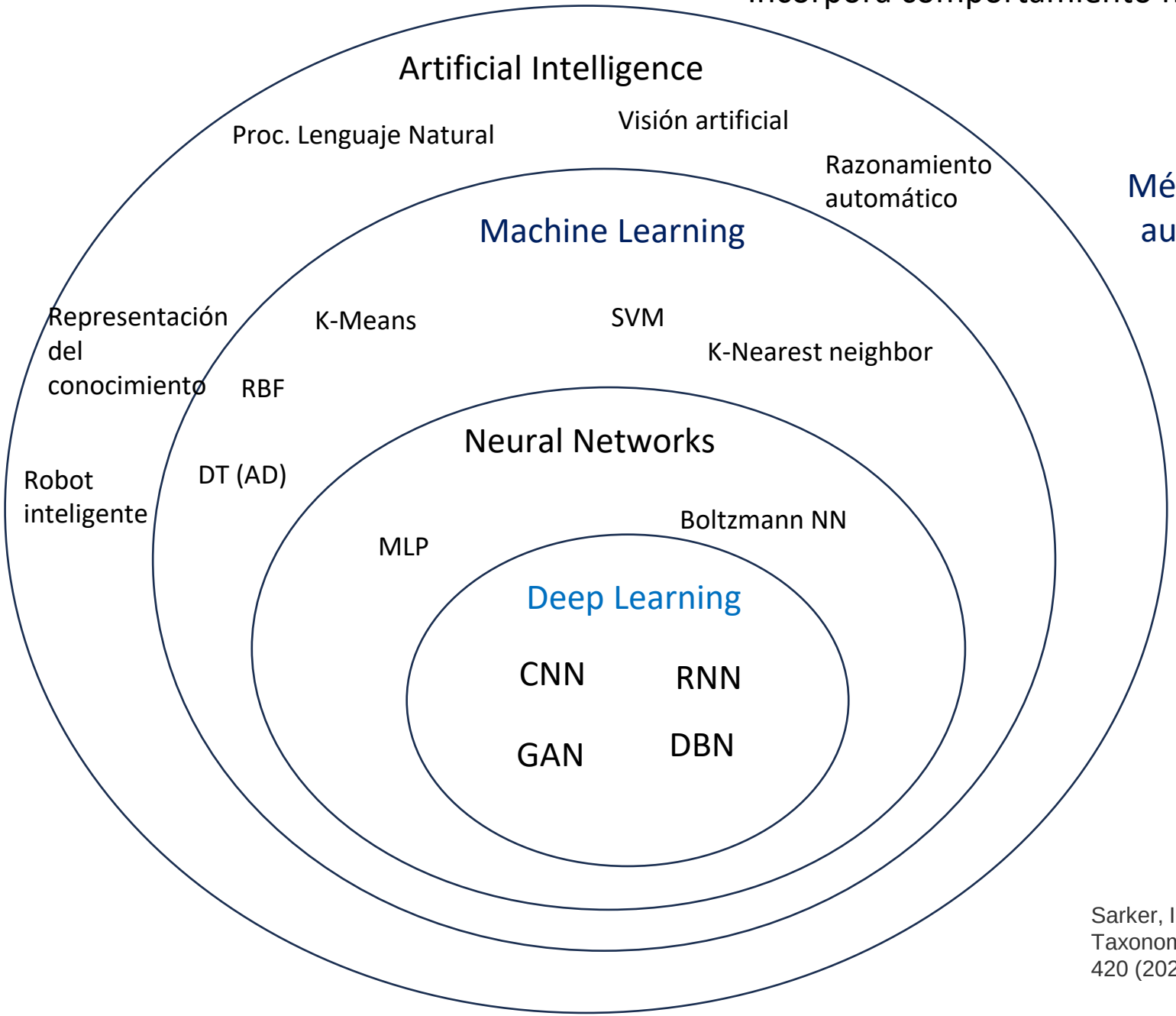
Módulo 3. Análisis Exploratorio de Datos

Equipo Docente:

Dra. Sonia I. Mariño
Lic. Lucia del Valle Ledesma
Lic. Rafael Perez

2024

Incorpora comportamiento humano e inteligencia a maquinas o sistemas



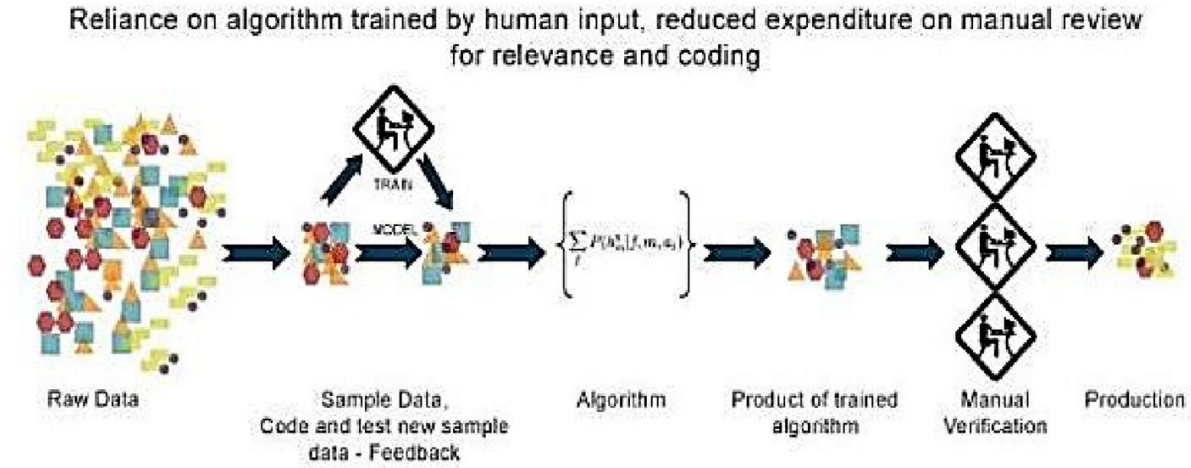
Métodos que aprenden de datos o experiencia, automatización en la construcción de modelos

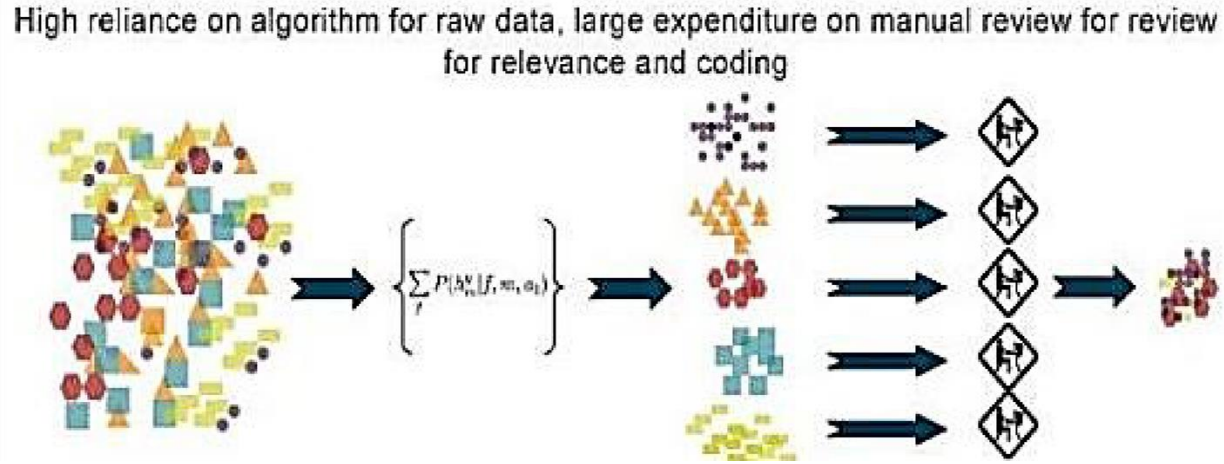
- Multilayer Perceptrons (MLPs)
- Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs)
- Long Short Term Memory Networks (LSTMs)
- Generative Adversarial Networks (GANs)
- Radial Basis Function Networks (RBFNs)
- Self Organizing Maps (SOMs)
- Deep Belief Networks (DBNs)
- Restricted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

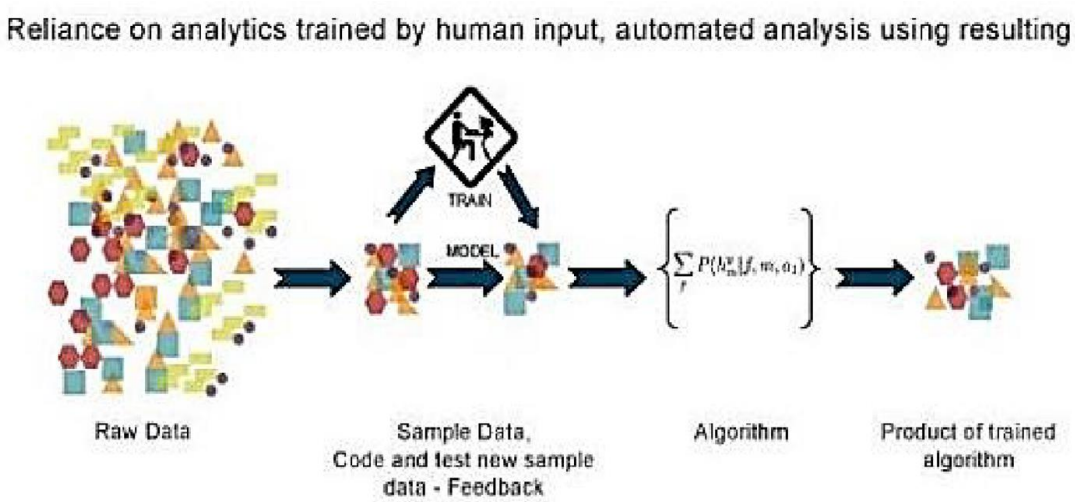
Computación utilizando RN multi-layer

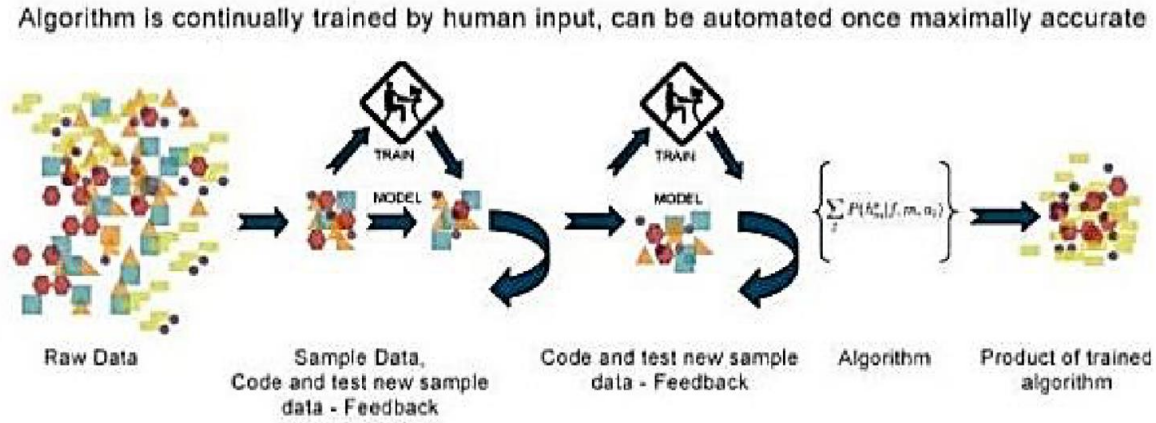
Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN COMPUT. SCI.* **2**, 420 (2021). https://doi.org/10.1007/s42979-021-00815-1

Aprendizaje automatico a partir de datos [conocimiento]

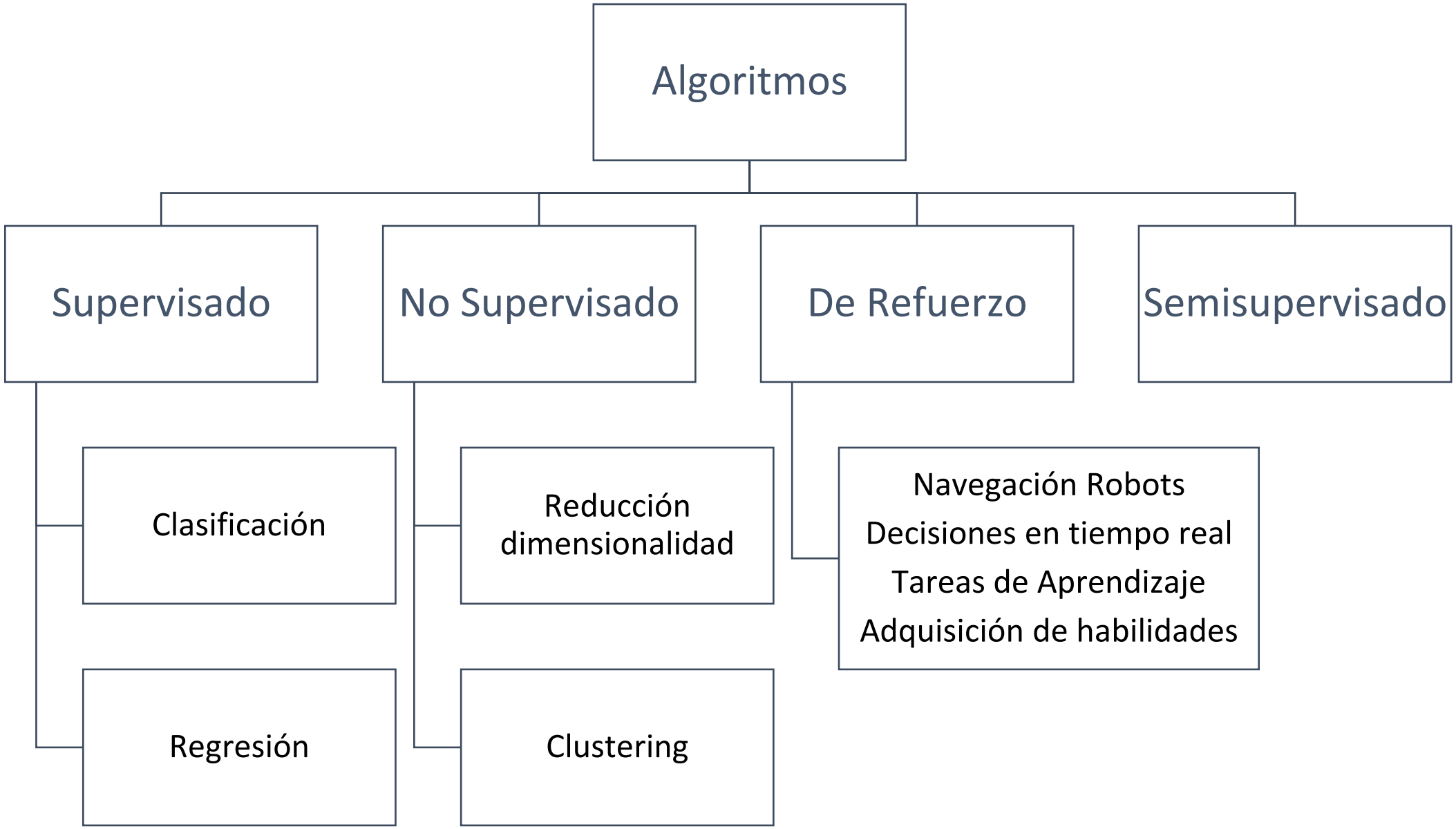








Refuerzo, retroalimentación


Algoritmos
Supervisado
No Supervisado
De Refuerzo
Semisupervisado
Clasificación
Regresión
Reducción dimensionalidad
Clustering
Navegación Robots
Decisiones en tiempo real
Tareas de Aprendizaje
Adquisición de habilidades

Proyectos ML / CD

- Perfiles profesionales
- Algunas definiciones un Proyecto ML / CD
- Metodologías, métodos, pipeline

Perfiles profesionales

Perspectivas

- RRHH con habilidades diferenciadas en Ingeniería del Software y ML / CD. IS carece de experiencia de ML / CD y comprende vocabulario de sus colegas analistas de datos.
- RRHH con habilidades en ML / CD e Ingeniería del Software.

Científico de datos

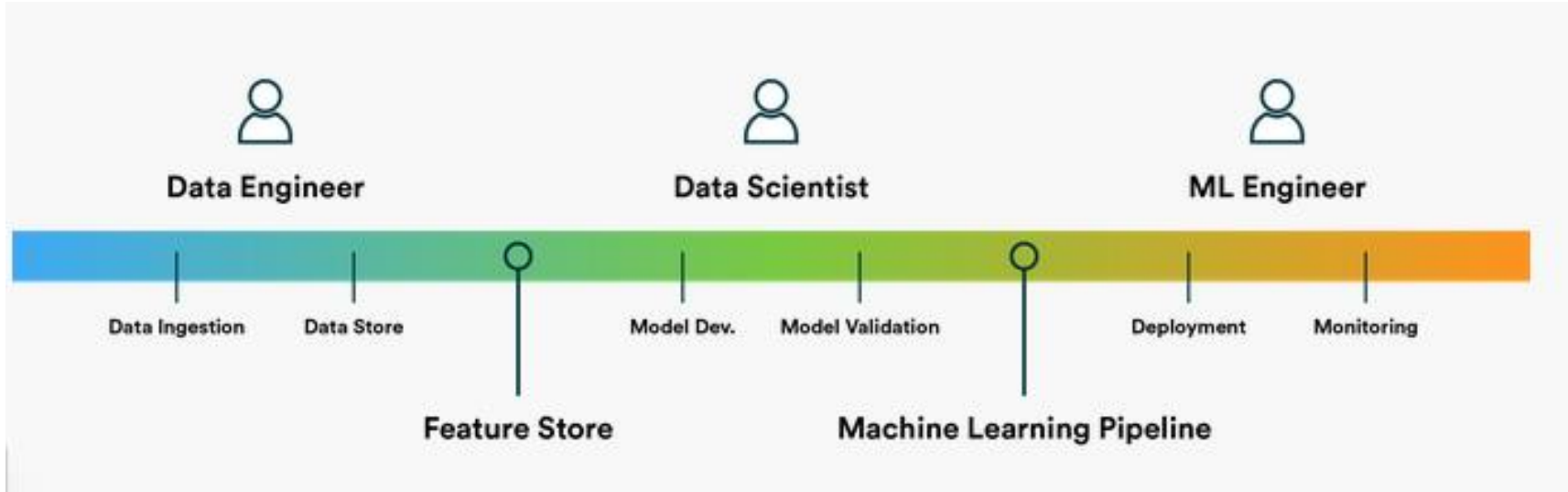
- Trata los datos recopilados, competencias para analizar requerimientos de los clientes, tratar los datos y agregar valor.
- Apoya toma de decisiones y la elaboración de una estrategia coherente y razonable.

Analista de datos

- Se especializa en un área de la CD.

Ingeniero de datos

- Responsable del funcionamiento del mantenimiento y del desarrollo de la infraestructura y de los sistemas de tratamiento de datos.



https://valohai.com/blog/the-three-roles-in-an-ml-team/

Cámara de la Industria Argentina del Software - CESSI

https://cessi.org.ar/perfiles-it/

- Desarrollo de Software
- **Análisis**
- Diseño
- Calidad de Software
- Soporte
- Implementadores
- Comunicación Online
- Seguridad

Analista Big Data (Data Scientist)

Es el responsable de interpretar y realizar descubrimientos en base a grandes volúmenes de información.

Nombres de cargos similares
Experto en Big Data, Analista Data Scientist, 'Chief data officer' (CDO), Analista Digital.

Consultor BI – Business Intelligence

Comprender y analizar el contexto de negocio y procesos de las organizaciones y en base a eso diseñar e implementar mejoras. Brindar soporte a la operación.

Nombres de cargos similares
Analista Business Intelligence,
Especialista en Business Intelligence.

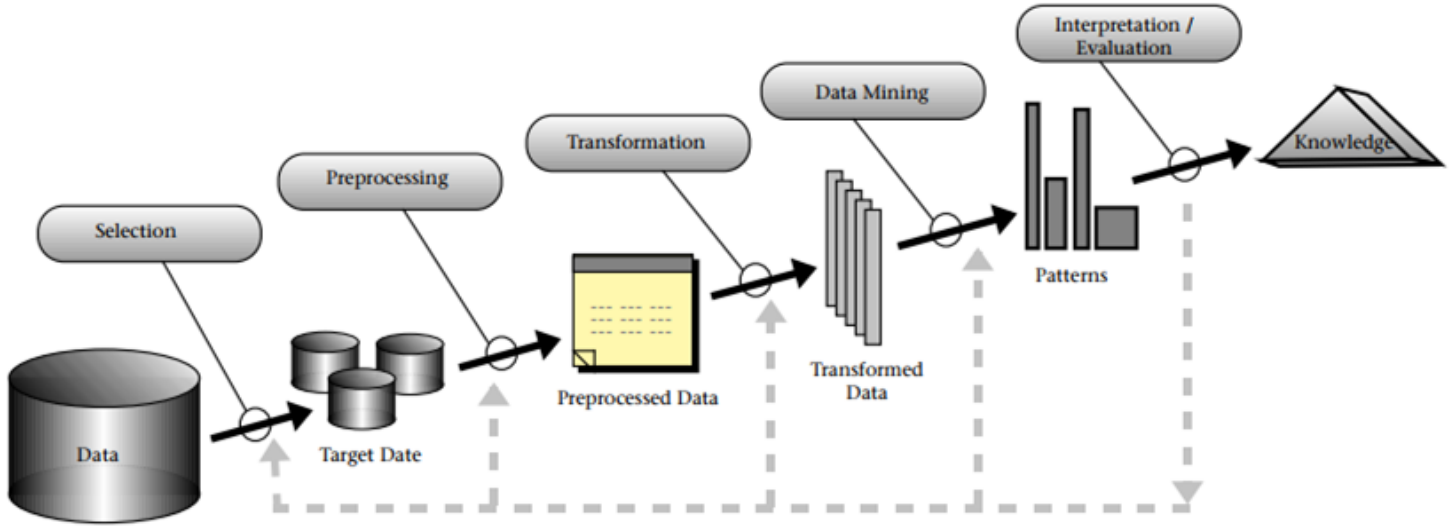
Especialistas del dominio

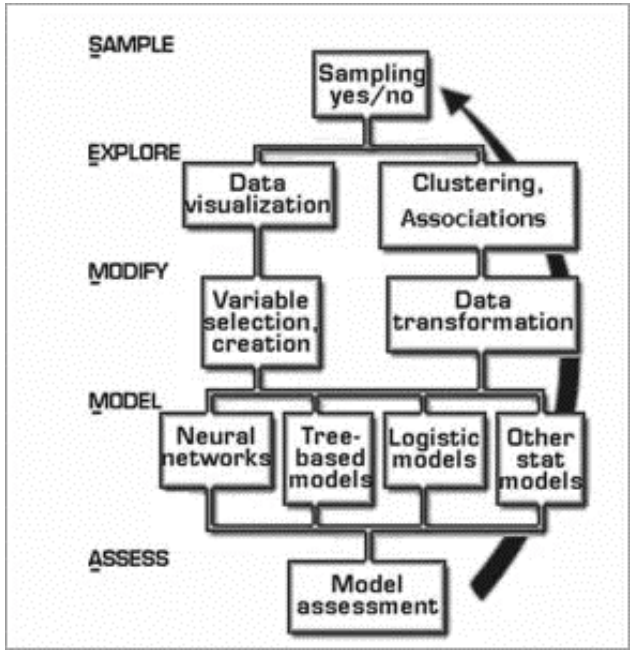
- Especialistas del dominio
  - Recomendable intervención, toma de decisiones sobre las entradas, salidas y características de su modelo.
- ¿Qué debería resolver el modelo?.
- ¿Qué se busca en los datos para obtener los resultados asociados a un negocio específico?.
- ¿Cómo transformar problema organizacional o empresarial en un problema de ML / CD ?

Algunas definiciones para proyectos de ML / CD

- Talento experimentado
- Apoyo de los líderes
- Infraestructura de datos
- Etiquetado de datos
- Colaboración en organizaciones
- Alineación entre los equipos técnicos y comerciales.
- Viabilidad técnica

Metodología, métodos, Pipeline





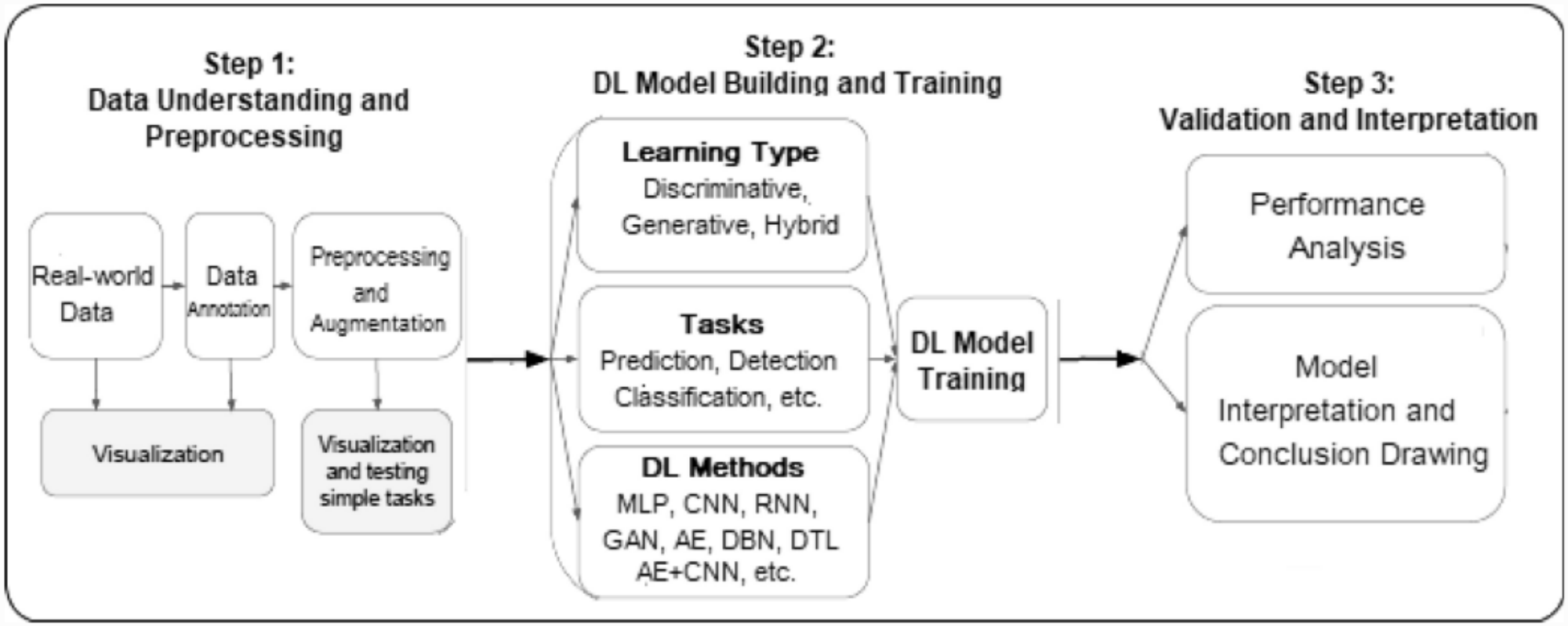
SEMMA, Sample, **Explore**, **Modify**, Model, and Assess.

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

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

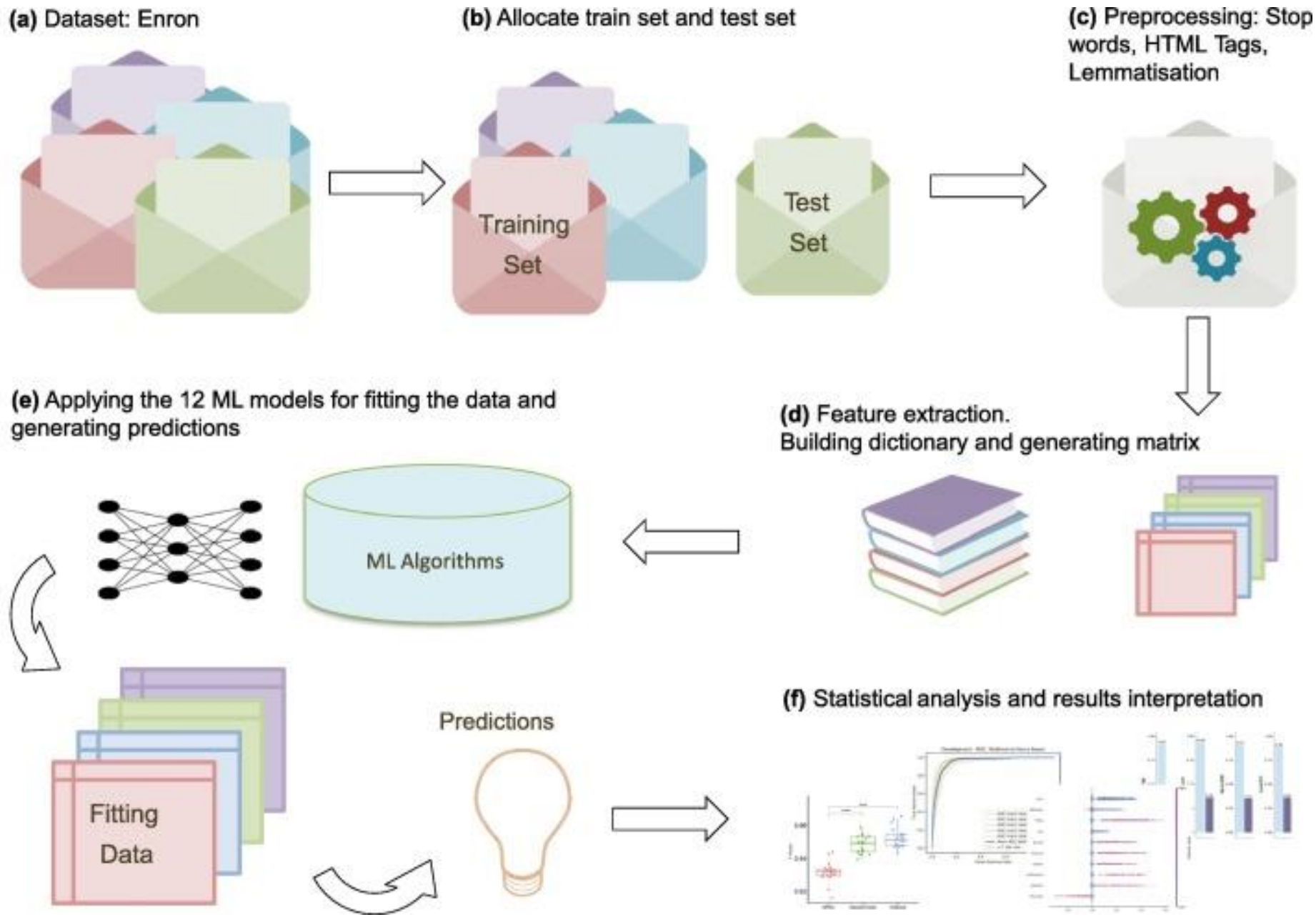
Fig. 4

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

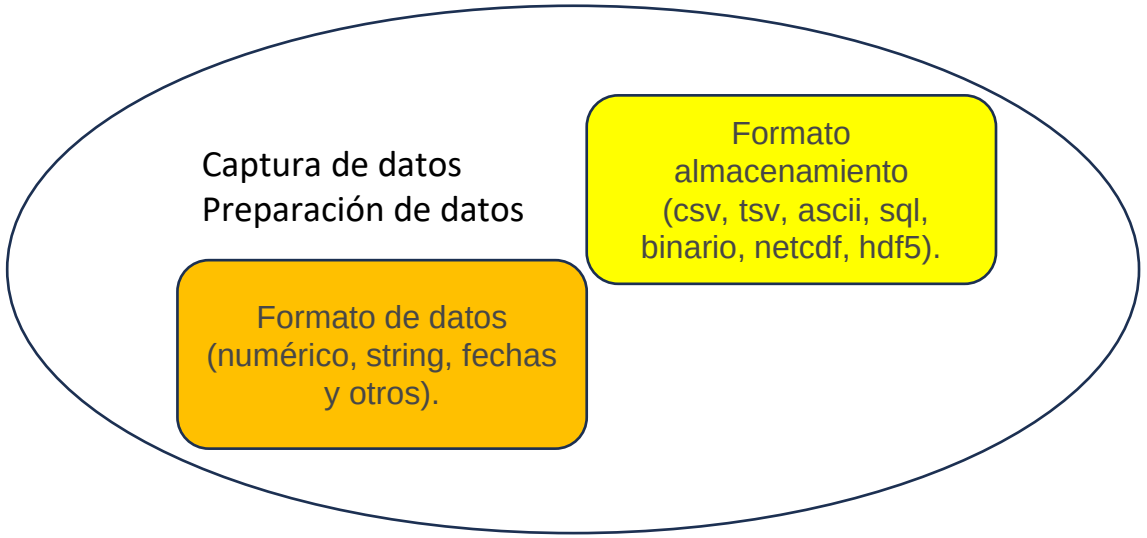
Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN COMPUT. SCI.* **2**, 420 (2021). https://doi.org/10.1007/s42979-021-00815-1



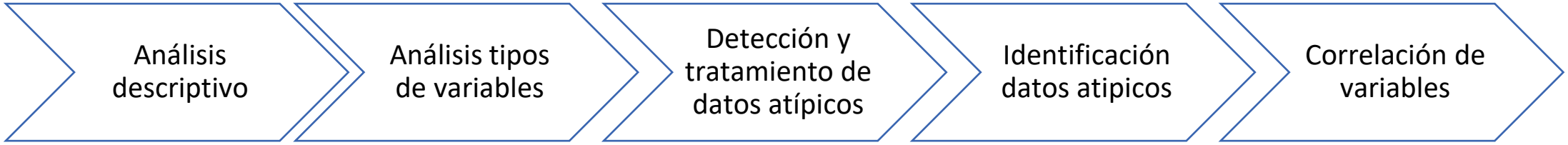
We proposed and tested a pipeline to compare and explain the classification outcomes of 12 machine learning models. We applied the pipeline for optimising and testing the models in a spam filtering context, with lemmatisation and noise-reduction techniques as preprocessing steps. The pipeline, which 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)

Proceso EDA



ANALISIS EXPLORATORIO DE DATOS



Documentar las decisiones en el proceso

Introducción

Análisis Exploratorio de Datos, EDA o Exploratory Data Analysis

- Refiere a una de las primeras tareas que desempeña el Científico de Datos.
- finalidad: explorar los datos de forma preliminar a la aplicación de cualquier proceso para una investigación o una visualización de datos.
- proceso fundamental para comprender los datos, identificar relaciones existentes, generar conocimientos, En contextos empíricos,
    - dado un conjunto de datos, se deberán generar respuestas en función a un objetivo de negocio y sus preguntas
    - requiere comprender el objetivo, vislumbrar posibles patrones, útiles para generar conocimiento a partir de datos

Introducción

EDA, relevancia de esta etapa en un proceso de CD,

- asegurar calidad en la información producida con tecnologías, fiabilidad y claridad para apoyar proceso de toma de decisiones.
- generar preguntas acerca de los datos.
- buscar respuestas visualizando, transformando y modelando los datos.
- aprender y refinar / mejorar las preguntas con la finalidad de diseñar nuevos interrogantes.

Introducción.

EDA,

ciclo iterativo, consiste en visualizar, transformar y modelar los datos

- organizar y preparar los datos,
- detectar fallos en el diseño y la captura de los datos,
- tratar y evaluar los datos ausentes, sesgos
- reconocer datos actuales
- identificar los casos atípicos y la posible relación que puedan existir entre las variables.

Proceso EDA.

Algunas buenas prácticas sobre los datos:

- Reproducibilidad
- Primero: los datos, después: el algoritmo

Método

Wickman y Grolemund (2017) propone como pasos para desarrollar un EDA:

- 1. Análisis **descriptivo de las variables**, para obtener una idea representativa del conjunto de datos.
- 2. **Ajuste de los tipos de las variables** para que sean consistentes en el momento de realizar posteriores operaciones.
- 3. **Detección y tratamiento de datos ausentes**, su tratamiento o eliminación es esencial para procesar adecuadamente las variables numéricas.
- 4. **Identificación y tratamiento de datos atípicos,** dado que pueden distorsionar futuros análisis estadísticos.
- 5. Examen numérico y gráfico de las **relaciones entre las variables analizadas para determinar el grado de correlación entre ellas**, predicción del comportamiento de una variable en función de otras.

EDA, niveles

Nivel 1 – descriptivo, univariado

- Centrado en el valor de un solo indicador

Nivel 2 – inferencial, bivariado

- Estudia una variable en función de otra.

Nivel 3 – modelización, multivariado

- Centrado en los indicadores disponible para estudiar un fenómeno determinado.
- Clasificación cruzada, análisis de varianza y regresiones simples.

Ingeniería de características

Ingeniería de características

- En ML / CD las características o entradas de un modelo se utilizan para su entrenamiento e inferencia
- Proceso de extraer y transformar las variables desde los datos sin procesar, en características que representen mejor el problema subyacente. Consiste en:
  - Crear funciones
  - Transformación de características
  - Extracción de características
  - Selección de características
- Aporta en la construcción de un modelo más preciso
- Ej. Agregar una característica “genero” en el archivo “ventas de casas” para determinar el genero del corredor / vendedor que logró mayores ventas.

Ingeniería de características

Técnicas comunes, y sus variantes más utilizadas

- manipular datos faltantes: eliminación de variables, imputación de media o mediana, valor más común
- manipular valores continuos: normalización Min-Max, estandarización
- manipular datos categóricos: codificación de etiqueta, valores dummy
- selección de características

Herramientas software

Phyton, librerías

NumPy:

- biblioteca para el procesamiento numérico.
- brinda estructuras de datos eficientes para el manejo de arreglos multidimensionales y operaciones matemáticas de alto rendimiento.

Pandas:

- biblioteca que proporciona estructuras de datos y herramientas para el análisis de datos.
- manipular y analizar datos tabulares, aplicable para preprocesamiento de datos antes de realizar análisis más complejos.

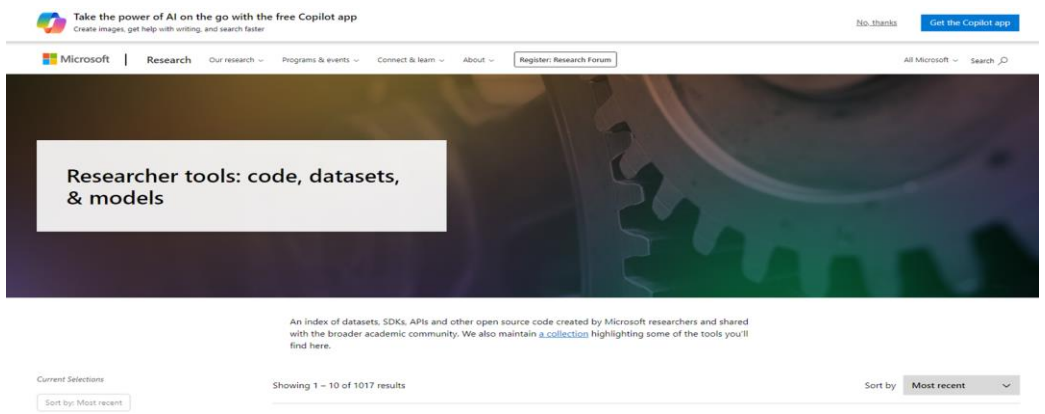
Matplotlib:

- biblioteca de visualización de datos en Python.
- herramientas para crear gráficos estáticos, gráficos interactivos y todo tipo de visualizaciones para explorar y comunicar resultados de análisis de datos.

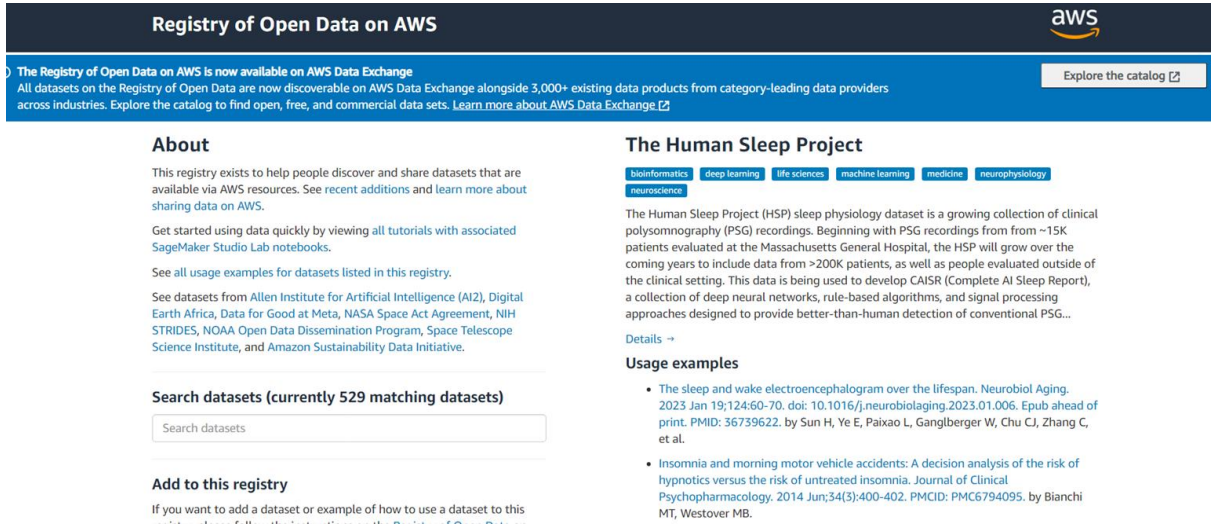
Seaborn

- biblioteca de visualización de datos en Python. Basada en Matplotlib

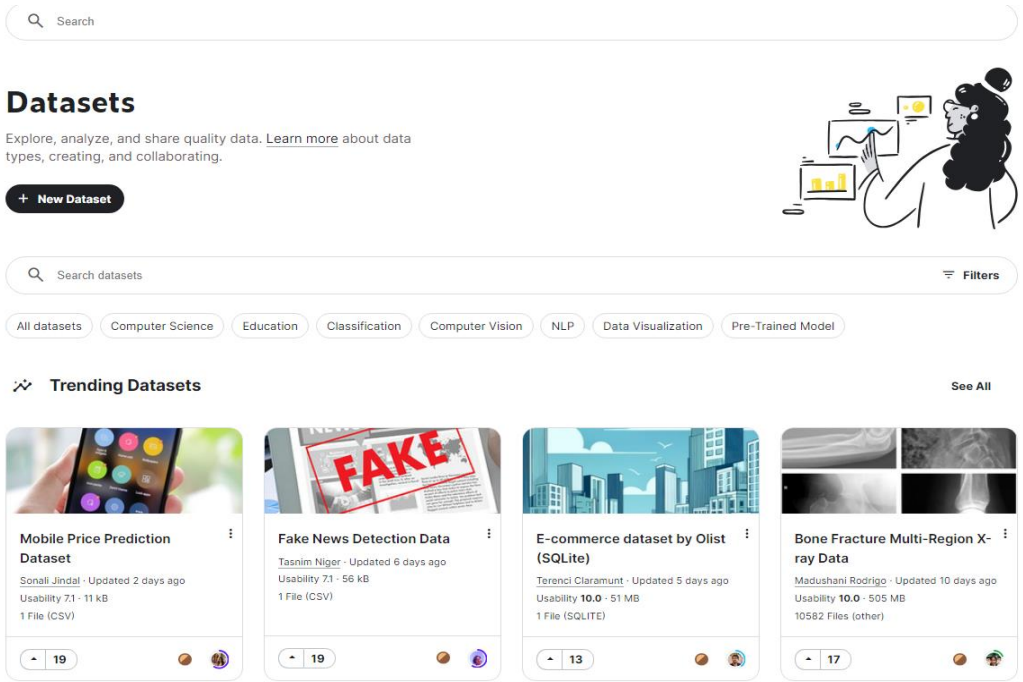
Recursos. Datasets



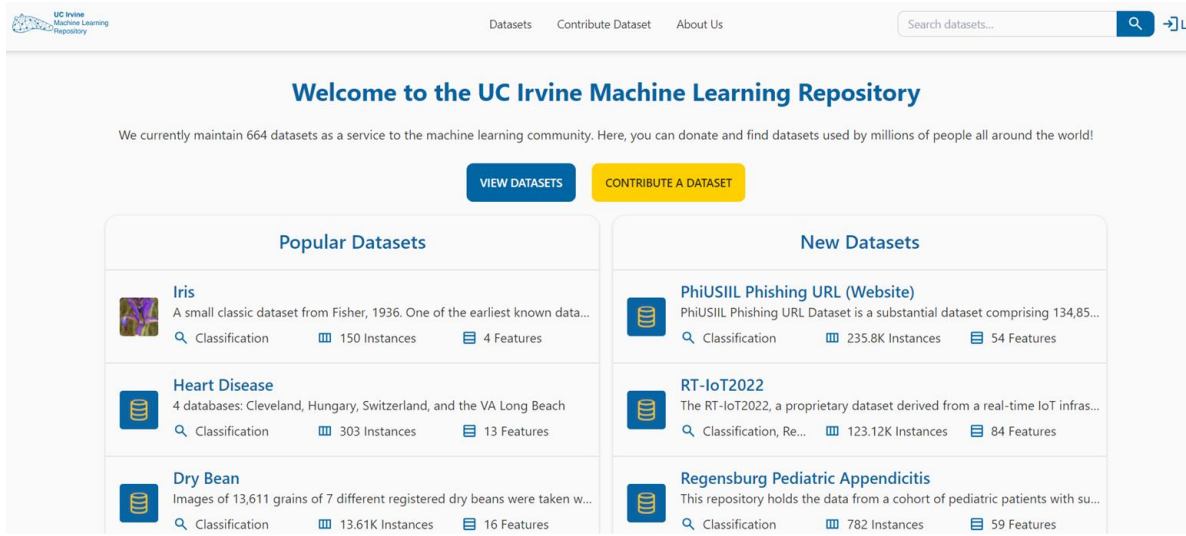
Researcher tools: code, datasets, & models - Microsoft Research



Registry of Open Data on AWS



https://www.kaggle.com/datasets



Home - UCI Machine Learning Repository

Tipología de datos

En EDA

- Datos estructurados y no estructurados.
- Formato de almacenamiento de datos (csv, tsv, ascii, sql, binario, netcdf, hdf5).
- Encabezados.
- Formato de los datos (numérico, string, fechas y otros).

Datos estructurados vs. no estructurados

Generalmente resultan en:

- Datos estructurados
- Datos no estructurados
- Datos semi estructurados

Datos estructurados

- Se organizan en un formato fijo y predefinido.
- Disponibles en archivos tipo texto, base de datos, hojas de cálculo. Presentados en columnas y filas
- Los títulos de las columnas están etiquetados
- Favorecen el procesamiento fácil utilizando herramientas software apropiadas
- Información ordenada, disponible en la mayoría de las bases de datos.
- Facilidad de gestión, digital y manual.
- Reduce los riesgos en el análisis con tecnologías de IA
- Algunos formatos
    - Hojas de cálculo o tablas de Excel
    - Bases de datos
    - Formularios web / formularios tipo test
    - Fichas estandarizadas de clientes
    - Encuestas a usuarios

Datos no estructurados

Datos binarios sin estructura comprensible para el lenguaje humano.

Objetos con datos sin organización. Su valor se obtiene al identificar y almacenar organizadamente

Organizar los elementos que conforman su contenido facilita su categorización para generar información.

Mayor dificultad para su procesamiento respecto a los datos estructurados.

No prevé su almacenamiento en herramienta de gestión de datos tradicional, debido a que no se incluyen en filas y columnas.

Datos no estructurados disponibles en

- Documentos en archivos de texto
- Imágenes
- Archivos PDF
- Archivos de registro y de datos de aplicaciones como .ini o .dll
- Datos de redes sociales: Instagram, Facebook y Twitter
- Vídeos
- Correos electrónicos
- Datos de ubicación
- Archivos de audio o grabaciones telefónicas

Datos semi estructurados

- Datos semi estructurados presentan un nivel medio de organización y clasificación
- Para agruparse y almacenarse utilizan los metadatos.
- Automatizarlos es más difícil -respecto a los **datos estructurados-,**
- Ejemplo de datos semi estructurados, disponibles en un servidor que almacena
  - datos no estructurados disponibles en los correos electrónicos
  - datos estructurados, en documentos adjuntos en los correos electrónicos

Utilidad de las estructuras de datos

Por que es recomendable el uso de estructuras de datos:

- Datos eficientes para el procesamiento
- Mejor gestión de la memoria
- Código más organizado

Formato de almacenamiento de datos

CSV (Comma-Separated Values):

- Comúnmente usado para datos tabulares, archivos de texto con datos separados por coma.

TSV (Tab-Separated Values):

- Similar al CSV, usa tabulaciones como delimitadores.

Excel

- Formato de planilla de calculo

JSON (JavaScript Object Notation):

- Formato de intercambio de datos que es fácil de leer y escribir para humanos y máquinas.

SQL (Structured Query Language):

- Usado para gestionar y consultar bases de datos relacionales.

Formato de almacenamiento de datos

Binario:

- Almacena datos en formato binario, eficiente para almacenamiento y transmisión. No es legible por humanos.
- Útil para almacenar datos complejos como imágenes, videos y archivos de aplicaciones

ASCII (American Standard Code for Information Interchange)

- conjunto de códigos que representan texto en computadoras y otros dispositivos que utilizan texto.
- archivos ASCII son archivos de texto plano, contienen datos codificados en el estándar ASCII.
- utilizados principalmente para almacenar texto sin formato y datos sencillos.

NetCDF (Network Common Data Form)

- formato de archivo diseñado para almacenar grandes volúmenes de datos científicos multidimensionales, como datos climáticos y meteorológicos.
- Los archivos NetCDF permiten la creación, acceso y intercambio de datos en un formato portátil y eficiente.

HDF5 (Hierarchical Data Format versión)

- formato de archivo, para almacenar y organizar grandes cantidades de datos complejos.
- aplicaciones científicas e ingeniería, almacena datos jerárquicamente y soporta datos multidimensionales y heterogéneos.

Introducción. Datos

- Formato de los datos
  - **numérico**, almacena números que pueden ser decimales o enteros.
  - **carácter**: alberga cadenas de texto.
  - **categórico**: contiene un número limitado de valores o categorías de información.
  - **lógico o booleano**: variables binarias, dos valores posibles: TRUE y FALSE ó 0 y 1; pueden ser resultado de una comparación o condición de otras variables presentes en el conjunto de datos.
  - **fecha**: almacena intervalos específicos de tiempo.