

05 MLAR-Aprendizaje supervisado

Tema 1 – Aprendizaje Automático

Curso Abril 23/24



Universidad
Internacional
de Valencia

¿Dónde estudiar el Tema 1?

Manual de la Asignatura – Tema 1

Scripts del Tema 1

03. Materiales del profesor > Recursos adicionales > Algoritmos > Algoritmo CRISP-ML(Q)

Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

Índice

1.1. Metodología

- 1.1.1. Introducción al aprendizaje automático
- 1.1.2. La metodología CRISP-DM
- 1.1.3. Roles en ciencia de datos
- 1.1.4. Comprensión del negocio
- 1.1.5. Comprensión de los datos
- 1.1.6. Preparación de los datos
- 1.1.7. Modelado
- 1.1.8. Evaluación de negocio
- 1.1.9. Despliegue

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

Índice

1.1. Metodología

1.1.1. Introducción al aprendizaje automático

1.1.2. La metodología CRISP-DM

1.1.3. Roles en ciencia de datos

1.1.4. Comprensión del negocio

1.1.5. Comprensión de los datos

1.1.6. Preparación de los datos

1.1.7. Modelado

1.1.8. Evaluación de negocio

1.1.9. Despliegue

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.1.1. Introducción al aprendizaje automático

El aprendizaje supervisado es un tipo de aprendizaje concreto dentro de lo que se conoce como **aprendizaje automático** (*machine learning* en inglés).

¿Qué es el *machine learning*?

- “ ... rama de conocimiento en la que se abordan algoritmos que son capaces de crear **modelos de conocimiento** abstractos a partir de históricos de **datos ... naturaleza empírica ...** ”

El aprendizaje automático es una rama de conocimiento que pertenece al campo de la **minería de datos** (*data mining* en inglés)

¿Qué es el *data mining*?

- “ ... conjunto de **procesos, métodos y técnicas** que conducen a la extracción de conocimiento a partir de bases de datos ... ”

Índice

1.1. Metodología

1.1.1. Introducción al aprendizaje automático

1.1.2. La metodología CRISP-DM

1.1.3. Roles en ciencia de datos

1.1.4. Comprensión del negocio

1.1.5. Comprensión de los datos

1.1.6. Preparación de los datos

1.1.7. Modelado

1.1.8. Evaluación de negocio

1.1.9. Despliegue

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.1.2. La metodología CRIPS-DM

¿Cómo llevar a cabo un proceso de DM/ML de forma racional y estructurada?

- **KDD** (*Knowledge Discovery from Databases*)... especialmente en el ámbito académico.
- **CRISP-DM** (Cross-Industry Standard Process for Data Mining) ... la más popular ... sí tienen en cuenta la proyección industrial de los proyectos
- **ASUM-DM**
- **CRISP-ML** (2021)
- **MLOps** (<https://ml-ops.org/>)

1.1.2. La metodología CRIPS-DM

CRISP-DM vs. CRISP-ML

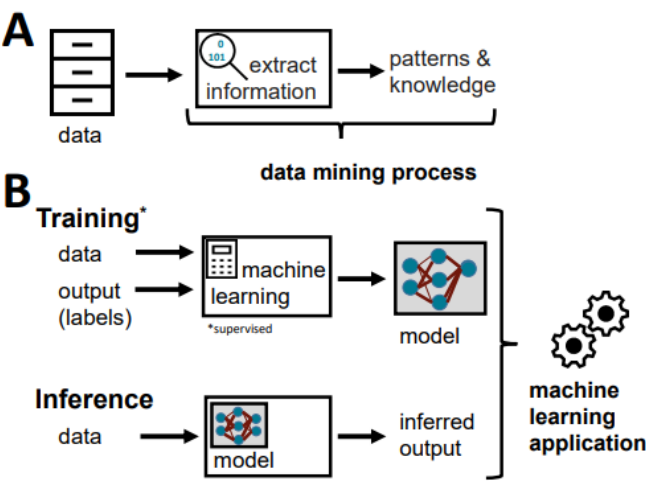


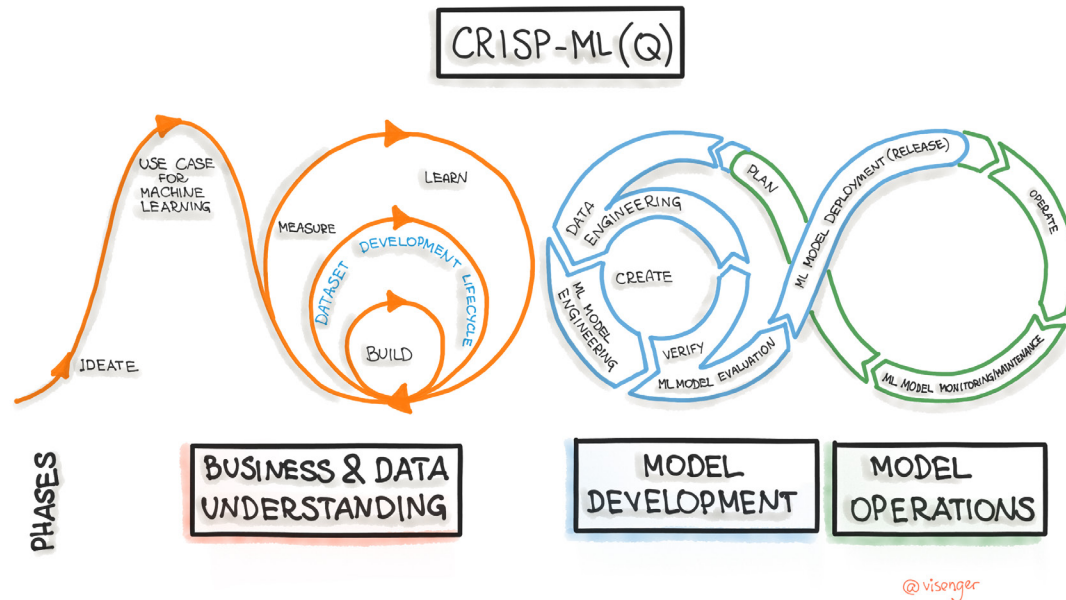
Figure 1. Difference between (A) data mining processes and (B) machine learning applications.

CRISP-ML(Q)	CRISP-DM	Amershi et al. [34]	Breck et al. [35]
Business and Data Understanding	Business Understanding	Requirements	-
Data Preparation	Data Preparation	Collection Cleaning Labeling Feature Engineering	Data
Modeling Evaluation Deployment Monitoring & Maintenance	Modeling Evaluation Deployment	Training Evaluation Deployment Monitoring	Model - - Monitoring

Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology (2021)

1.1.2. La metodología CRISP-DM

CRISP-ML



<https://ml-ops.org/content/crisp-ml>

1. Business and Data Understanding
2. Data Engineering (Data Preparation)
3. Machine Learning Model Engineering
4. Quality Assurance for Machine Learning Applications
5. Deployment
6. Monitoring and Maintenance

Índice

1.1. Metodología

1.1.1. Introducción al aprendizaje automático

1.1.2. La metodología CRISP-DM

1.1.3. Roles en ciencia de datos

1.1.4. Comprensión del negocio

1.1.5. Comprensión de los datos

1.1.6. Preparación de los datos

1.1.7. Modelado

1.1.8. Evaluación de negocio

1.1.9. Despliegue

1.2. Tipos de aprendizaje

1.3. Estructura de datos

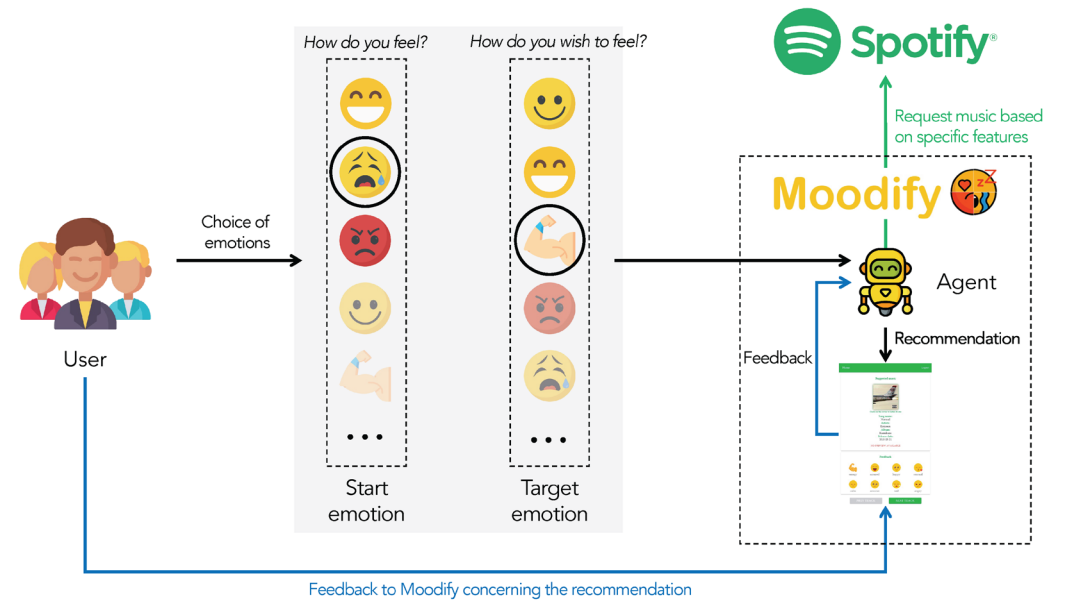
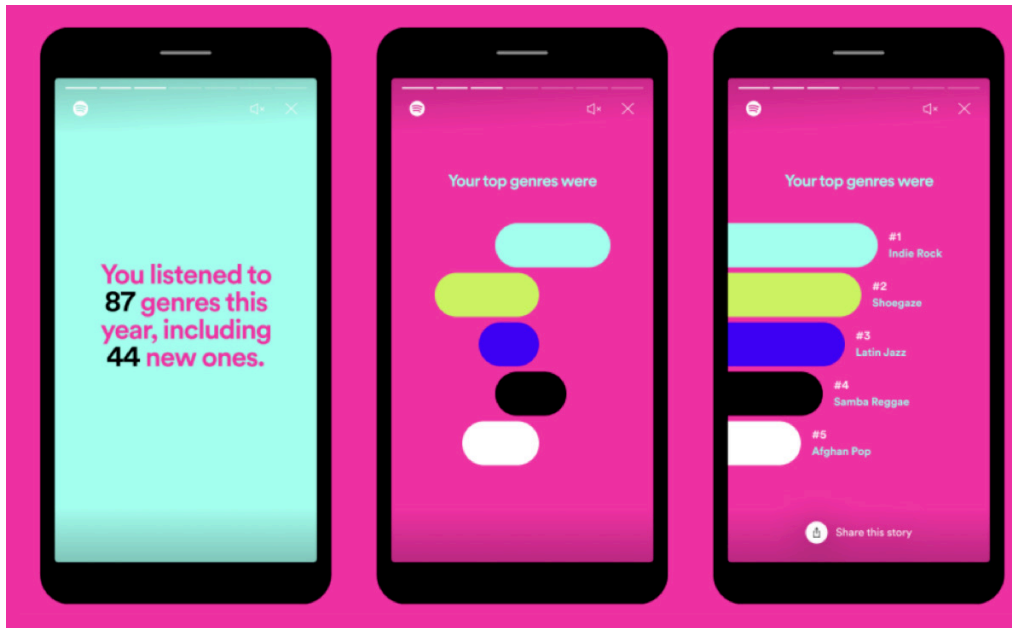
1.4. Limpieza de datos

1.1.3. Roles en ciencia de datos

En el mundo real, los roles pueden estar solapados. Depende del nivel organizativo de la empresa/organización.

- **Director técnico de proyecto** (*data scientist chief*): “**planifica el proyecto, asigna recursos, supervisa el trabajo del equipo**, establece prioridades y realiza la **validación última de los resultados del proyecto.**”
- **Científico de datos** (*data scientist*): “**analizar los problemas, comunicarse con expertos y personal técnico**, diseñar e implementar las **soluciones** más apropiadas, **diseñar** y ejecutar la **experimentación** e **interpretar los resultados.**”
- **Analista de datos** (*data analyst*): “análisis de datos e inteligencia de negocio (**business intelligence** en inglés) con el objetivo de **extraer conclusiones útiles a partir de los datos y elaborar informes y cuadros de mandos**” <https://powerbi.microsoft.com/en-us/>
- **Ingeniero de datos** (*data engineer*): “maneja los datos con destreza utilizando herramientas de transformación de datos, sistemas gestores de bases de datos y lenguajes de programación para extraer, limpiar, transformar y cargar los datos de forma estructurada”

1.1.3. Roles en ciencia de datos



<https://www.kdnuggets.com/2015/11/different-data-science-roles-industry.html>

Índice

1.1. Metodología

1.1.1. Introducción al aprendizaje automático

1.1.2. La metodología CRISP-DM

1.1.3. Roles en ciencia de datos

1.1.4. Comprensión del negocio

1.1.5. Comprensión de los datos

1.1.6. Preparación de los datos

1.1.7. Modelado

1.1.8. Evaluación de negocio

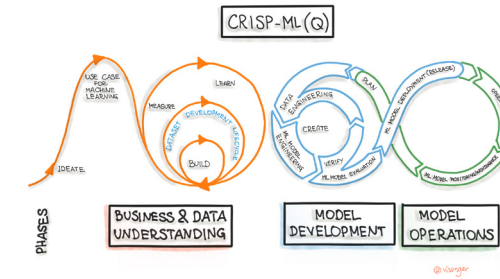
1.1.9. Despliegue

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.1.4. Comprensión del negocio



“En esta fase se establecen los requisitos y objetivos del proyecto desde una perspectiva empresarial para luego trasladarlos a objetivos técnicos y a un plan de proyecto. Para llevar a cabo esta fase, es necesario comprender en profundidad el problema que se quiere resolver.”

- **Determinar los objetivos de negocio:** ¿Cuál es el problema a resolver? ¿Por qué necesitamos utilizar el ML/DM? ¿Qué se considera un proyecto exitoso?
- **Evaluar la situación actual:** ¿Cuál es la situación actual? ¿Qué solución tenemos actualmente implementada en nuestra organización? ¿Qué soluciones hay actualmente en el estado del arte? ¿Tenemos los recursos necesarios para afrontar el problema?
- **Determinar los objetivos de la minería de datos.**
- **Producir un plan de proyecto.**

Índice

1.1. Metodología

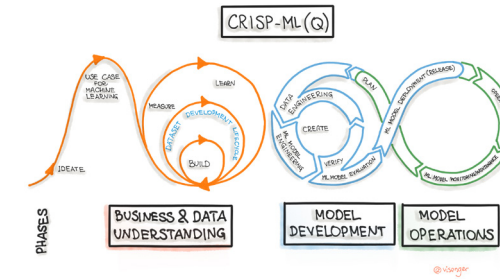
- 1.1.1. Introducción al aprendizaje automático
- 1.1.2. La metodología CRISP-DM
- 1.1.3. Roles en ciencia de datos
- 1.1.4. Comprensión del negocio
- 1.1.5. Comprensión de los datos**
- 1.1.6. Preparación de los datos
- 1.1.7. Modelado
- 1.1.8. Evaluación de negocio
- 1.1.9. Despliegue

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.1.5. Comprensión de los datos



“En esta fase se lleva a cabo la recolección y exploración inicial de los datos, con el objetivo de establecer un primer contacto con el problema. Esta fase suele ser crítica en el proyecto, pues el buen entendimiento de los datos tiene como consecuencia una importante reducción en el tiempo global del proyecto, además de incrementar las garantías de éxito. ”

- **Recolectar datos iniciales:** “Se deben elaborar informes que incluyan una lista de los datos adquiridos, su localización, las técnicas utilizadas...”
- **Describir los datos:** “... el número de conjuntos de datos, sus filas y columnas, su identificación, el significado de cada columna ...”
- **Explorar los datos:** “... técnicas básicas de estadística descriptiva que revelan propiedades de los datos ... ”
- **Verificar la calidad de los datos:** “...consistencia, la cantidad y distribución de los valores nulos o valores fuera de rango...”

Índice

1.1. Metodología

1.1.1. Introducción al aprendizaje automático

1.1.2. La metodología CRISP-DM

1.1.3. Roles en ciencia de datos

1.1.4. Comprensión del negocio

1.1.5. Comprensión de los datos

1.1.6. Preparación de los datos

1.1.7. Modelado

1.1.8. Evaluación de negocio

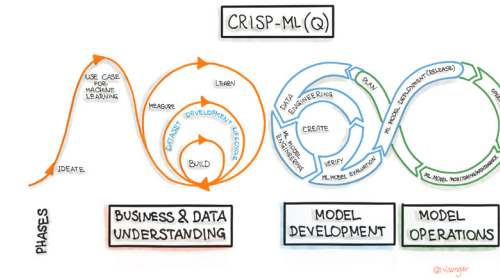
1.1.9. Despliegue

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.1.6. Preparación de los datos



“En esta fase se trata de seleccionar, limpiar y generar conjuntos de datos correctos, organizados y preparados para la fase de modelado. ”

- **Seleccionar los datos:** “... subconjunto de datos considerando la calidad y consistencia de los datos...”
- **Limpiar los datos:** “... mejorar la calidad y consistencia de los datos para la fase de modelado (1.1.7)... ”
- **Construir datos**
- **Integrar los datos:** “...crear conjuntos de datos fusionados y homogéneos...”
- **Formatear los datos:** “cumplir con ciertas interfaces funcionales, por ejemplo, con una interfaz de programación de aplicaciones”

Índice

1.1. Metodología

1.1.1. Introducción al aprendizaje automático

1.1.2. La metodología CRISP-DM

1.1.3. Roles en ciencia de datos

1.1.4. Comprensión del negocio

1.1.5. Comprensión de los datos

1.1.6. Preparación de los datos

1.1.7. Modelado

1.1.8. Evaluación de negocio

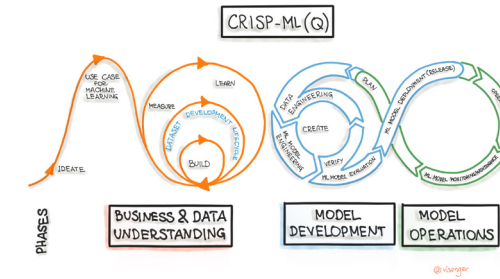
1.1.9. Despliegue

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.1.7. Modelado



“En esta tarea se aborda el problema de la **selección de los algoritmos más adecuados** para el **conjunto de datos de estudio**. Aunque los algoritmos más eficaces se descubren tras un **análisis comparativo** de resultados en conjuntos de **datos de prueba**, a veces es conveniente seleccionar un **subconjunto de algoritmos candidatos** y someterlos a prueba para reducir los tiempos de experimentación.”

- Selección de los algoritmos de modelado
- Generar el plan de prueba
- Construir los modelos
- Evaluación de resultados

Índice

1.1. Metodología

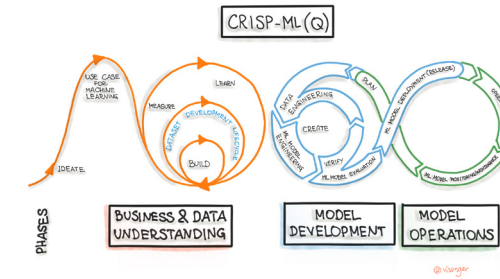
- 1.1.1. Introducción al aprendizaje automático
- 1.1.2. La metodología CRISP-DM
- 1.1.3. Roles en ciencia de datos
- 1.1.4. Comprensión del negocio
- 1.1.5. Comprensión de los datos
- 1.1.6. Preparación de los datos
- 1.1.7. Modelado
- 1.1.8. Evaluación de negocio**
- 1.1.9. Despliegue

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.1.8. Evaluación del negocio



“Se **evalúan los modelos obtenidos**, teniendo en cuenta el **cumplimiento de los criterios de éxito** del problema de negocio planteado al inicio del proyecto y usando el **conocimiento del dominio del problema**. Además, se deben resumir los **hallazgos o conclusiones** más importantes que se pueden extraer de los **resultados experimentales** obtenidos en **todas las fases anteriores**, tanto en la comprensión y preparación de datos como en la comprensión del problema y el modelado.”

- **Evaluar los resultados**
- **Revisar el proceso**
- **Determinar los siguientes pasos:** “ ... iniciar nuevas iteraciones en el proyecto o generar nuevos proyectos o subproyectos... ”

Índice

1.1. Metodología

1.1.1. Introducción al aprendizaje automático

1.1.2. La metodología CRISP-DM

1.1.3. Roles en ciencia de datos

1.1.4. Comprensión del negocio

1.1.5. Comprensión de los datos

1.1.6. Preparación de los datos

1.1.7. Modelado

1.1.8. Evaluación de negocio

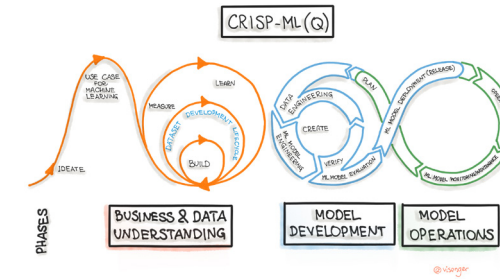
1.1.9. Despliegue

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.1.9. Despliegue



“En esta fase el conocimiento obtenido se transforma en **acciones dentro del proceso de negocio**, empleando los modelos construidos y los hallazgos obtenidos en la actividad productiva empresarial. Las tareas que se realizan **son planear la implementación, monitorizar y mantener, crear el informe final y revisar el proyecto.**”

- Planear la implementación
- Monitorizar y mantener
- Crear el informe final
- Revisar el proyecto

Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.1.1. Aprendizaje supervisado

1.1.2. Aprendizaje no supervisado

1.1.3. Otros tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.1.1. Aprendizaje supervisado

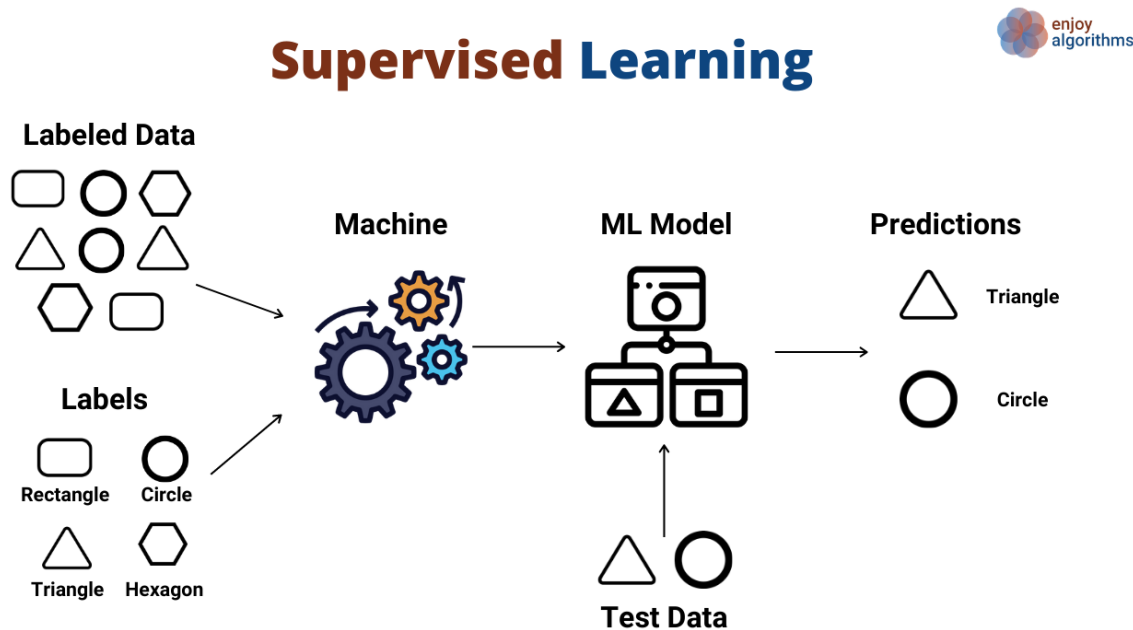
1.1.2. Aprendizaje no supervisado

1.1.3. Otros tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.2.1. Aprendizaje supervisado



- **Regresión.** Si la variable de salida es numérica (número real).
- **Clasificación.** Si la variable es categórica (número discreto o etiqueta nominal).

Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.1.1. Aprendizaje supervisado

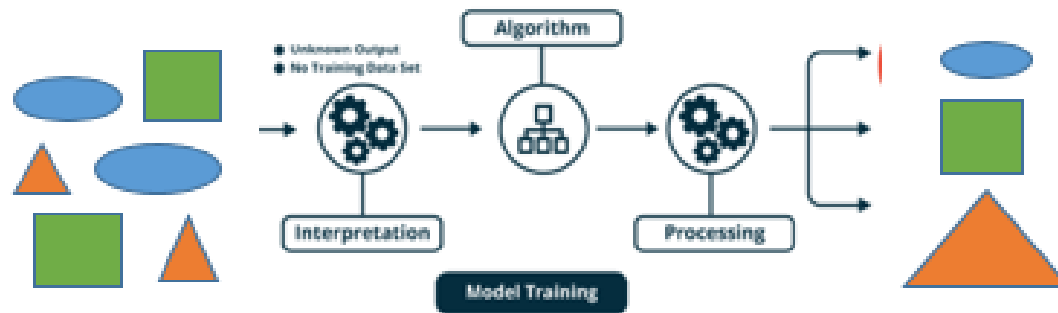
1.1.2. Aprendizaje no supervisado

1.1.3. Otros tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.2.1. Aprendizaje no supervisado



- **Clustering.** Creación de grupos de datos similares, de gran importancia para dividir los problemas, y analizar y modelar cada grupo de forma separada.
- **Reglas de asociación.** Creación de reglas que asocian las variables entre sí mediante relaciones causa-efecto (antecedente y consecuente), muy útil para estudiar las dependencias entre los datos y para los sistemas de recomendación.

Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.1.1. Aprendizaje supervisado

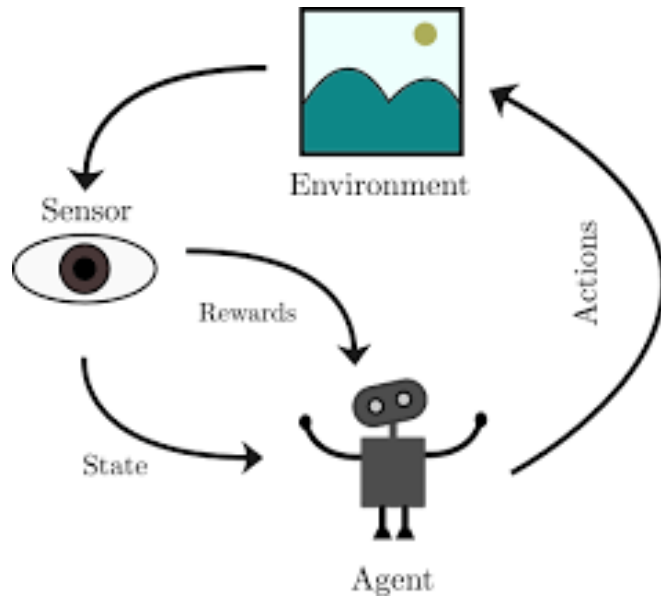
1.1.2. Aprendizaje no supervisado

1.1.3. Otros tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.2.1. Aprendizaje por refuerzo



- Modelado de relaciones agente-entorno con feedback.
- El aprendizaje viene dado por las recompensas.

Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.3.1. Estructuras y tipos en ciencia de datos

1.3.2. Instancias, clase, atributos, y modelos

1.1.3. Características de los conjuntos de datos

1.4. Limpieza de datos

Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.3.1. Estructuras y tipos en ciencia de datos

1.3.2. Instancias, clase, atributos, y modelos

1.1.3. Características de los conjuntos de datos

1.4. Limpieza de datos

1.3.1. Estructura y tipos en ciencia de datos

“ ... existe una estructura de datos fundamental en la que se apoyan la gran mayoría de algoritmos: la **tabla de datos**, también conocida como **conjunto de datos** (*dataset* en inglés). “

“La tabla de datos es una estructura en **dos dimensiones** (con filas y columnas), al igual que las matrices en matemáticas. No obstante, a diferencia de las matrices, las tablas de datos **pueden contener diferentes tipos de datos** (números, cadenas, fechas...). A pesar de ello, **cada columna de una tabla de datos debe contener solo datos del mismo tipo.** ”

Tipos de datos:

- **Numérico:** “45,6”/”49.1”
- **Categorico:** “Verde”/”Rojo”

Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.3.1. Estructuras y tipos en ciencia de datos

1.3.2. Instancias, clase, atributos, y modelos

1.1.3. Características de los conjuntos de datos

1.4. Limpieza de datos

1.3.2. Instancia, clase, atributos y modelo

- **Instancia:** ejemplo, muestra, observación, punto, o prototipo es cada fila de los datos.
- **Clase:** etiqueta, variable objetivo, respuesta, efecto, o consecuencia, es la variable de salida, objetivo del aprendizaje supervisado.
- **Atributo:** característica, causa, variable de entrada, variable explicativa, es cada una de las demás variables en la tabla de datos, excepto la clase.
- **Modelo:** cometido de un algoritmo de aprendizaje supervisado es la extracción inteligente de un modelo, a partir de las relaciones encontradas entre los atributos y la clase de las instancias de la tabla de datos.

Tabla 1

Estructura de datos basada en instancias, atributos y clase para minería de datos

		Atributos			Clase
		x_1	...	x_p	y
X	e_1	x_{11}	...	x_{1p}	y_1
	
	e_n	x_{n1}	...	x_{np}	y_n

Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.3.1. Estructuras y tipos en ciencia de datos

1.3.2. Instancias, clase, atributos, y modelos

1.1.3. Características de los conjuntos de datos

1.4. Limpieza de datos

1.3.3. Características de los conjuntos de datos

“Es habitual observar y anotar ciertas propiedades de los conjuntos de datos al inicio de los proyectos de minería de datos con el fin de catalogar o enfocar el tipo de problema que se pretende resolver.. “

Carácterísticas:

- **Número de instancias**
- **Número de atributos**
- **Nombre, tipo y breve descripción de cada atributo**
- **Tipo de datos de la clase, si tiene, y dominio de valores**
- **Nombre, y breve descripción de la clase, si tiene**
- **Cantidad de valores ausentes**

Tabla 2

Características del conjunto de datos "iris"

Número de instancias	150
Número de atributos	4
Descripción de atributos	sepal-length. Numérico. Longitud del sépalo. sepal-width. Numérico. Anchura del sépalo. petal-length. Numérico. Longitud del pétalo. petal-width. Numérico. Anchura del pétalo.
Tipo de datos de la clase	Categorico: {iris-setosa, iris-versicolor, iris-virginica}
Descripción de la clase	Nombre: class. Descripción: Especie de la planta.
Valores ausentes	0

https://scikit-learn.org/stable/datasets/toy_dataset.html

Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.4.1. Introducción a la limpieza de datos

1.4.2. Normálización y estandarización

1.4.3. Detección de outliers

1.4.4. Imputación de valores ausentes

1.4.5. Selección de atributos

Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.4.1. Introducción a la limpieza de datos

1.4.2. Normalización y estandarización

1.4.3. Detección de outliers

1.4.4. Imputación de valores ausentes

1.4.5. Selección de atributos

1.4.1. Introducción a la limpieza de datos

“La limpieza o preprocesado de datos consiste en la **eliminación de defectos en los datos y la adaptación de estos de cara a la fase posterior de modelado**, con el objetivo de **producir la forma de los datos que mejor exponga sus patrones**, de la forma más nítida posible... **pueden ser tanto supervisadas como no supervisadas ...**”

- **A nivel de muestra** (fila)
- **A nivel de variable** (columna)
- **A nivel de muestra-variable** (matriz)

“... estudiaremos algunas de las tareas típicas dentro de la limpieza de datos, tales como la **normalización y estandarización de atributos**, la **detección de outliers**, la **imputación de valores ausentes y la selección de atributos**. Existen, no obstante, **otras** numerosas tareas de limpieza de datos, como la **discretización**, la **generación de atributos o la selección de instancias**, entre otras ...”

Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.4.1. Introducción a la limpieza de datos

1.4.2. Normalización y estandarización

1.4.3. Detección de outliers

1.4.4. Imputación de valores ausentes

1.4.5. Selección de atributos

1.4.2. Normalización y estandarización

- **Normalización**

“La normalización de un conjunto de datos consiste en transformar los valores de sus atributos numéricos al rango continuo [0, 1] mediante una operación matemática llamada **homotecia**. Esta transformación se lleva a cabo para cada columna (atributo) de forma independiente. ...”

$$x_{ij}^N = \left\{ \frac{x_{ij} - \min(x_i)}{\max(x_i) - \min(x_i)}, \forall j \in 1, \dots, n \right\}$$

centrado

- **Estandarización**

“La estandarización consiste en transformar la distribución de los valores de los atributos para que estos posean una media y desviación típica determinadas. Esta transformación se lleva a cabo para cada atributo de forma independiente.”

$$x_{ij}^S = \left\{ \frac{x_{ij} - \text{media}(x_i)}{\text{desv}(x_i)}, \forall j \in 1, \dots, n \right\}$$

z-score

Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.4.1. Introducción a la limpieza de datos

1.4.2. Normalización y estandarización

1.4.3. Detección de outliers

1.4.4. Imputación de valores ausentes

1.4.5. Selección de atributos

1.4.3. Detección de outliers

“presencia de **valores** anómalos que **se encuentran fuera de la distribución natural de los datos.** ...”

“Determinar qué valores se pueden considerar outliers dentro de un conjunto de datos **no es un problema trivial**, pues **la frontera entre outliers y no outliers generalmente es difusa** y, aun así, es preciso determinar un umbral a partir del cual considerar cada muestra como outlier o no.”

- **Envolvente elíptica:** únicamente con datos de distribución normal. Trabaja a nivel de matriz.
- **Isolation Forest:** Trabaja a nivel de matriz.
- **Box-Whiskers:** Trabaja a nivel de atributo.
- ...

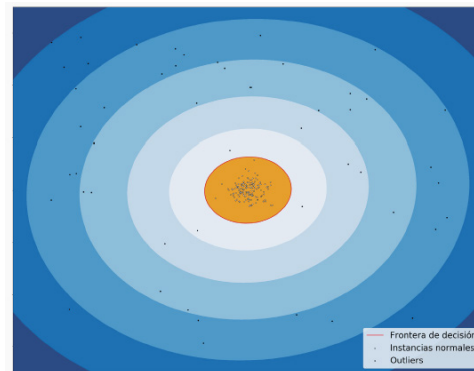
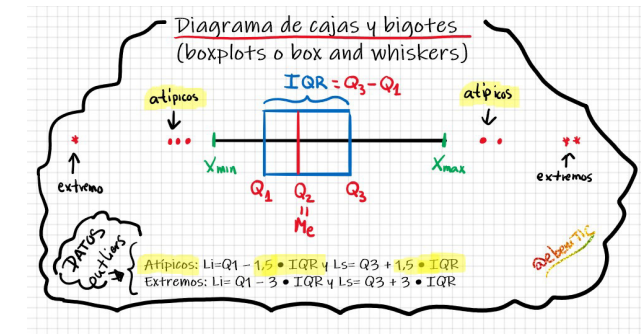
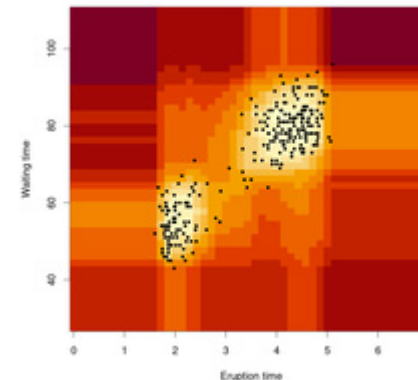


Figura 1. Gráfica de detección de outliers.



Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.4.1. Introducción a la limpieza de datos

1.4.2. Normalización y estandarización

1.4.3. Detección de outliers

1.4.4. Imputación de valores ausentes

1.4.5. Selección de atributos

1.4.4. Imputación de valores ausentes

“... los conjuntos de **datos** suelen estar **incompletos**, y varias de las **instancias** no poseen valores para todos sus **atributos**...”

“... Los datos que faltan son los valores ausentes cuyo valor en programación suele ser **NaN, NA, N/A o Null**...”

- **Eliminar instancias**
- **Eliminar muestras**
- **Imputar valores**
 - **Media**
 - **Mediana**
 - **Moda**
 - **Modelo destimación**

Índice

1.1. Metodología

1.2. Tipos de aprendizaje

1.3. Estructura de datos

1.4. Limpieza de datos

1.4.1. Introducción a la limpieza de datos

1.4.2. Normalización y estandarización

1.4.3. Detección de outliers

1.4.4. Imputación de valores ausentes

1.4.5. Selección de atributos

1.4.5. Selección de atributos

“... subtarea dentro de la tarea de reducción de dimensionalidad, la cual, a su vez, se incluye dentro de la limpieza de datos... supervisadas como no supervisadas ...”

“Todas las técnicas de selección de atributos necesitan de un **algoritmo de evaluación de atributos** que les permita determinar la bondad de estos con el fin de seleccionar aquellos que puedan ser candidatos para ser eliminados. ”

- **Evaluación de atributos individuales**
- **Evaluación de conjuntos de atributos:** “... la relevancia, o poder predictivo, que un atributo puede tener de cara a la predicción de una clase (problema de inferencia mediante aprendizaje supervisado) puede depender de la existencia de otros atributos... “ “... **algoritmos de búsqueda**”

Algunas estrategias:

- **Eliminación por alta varianza:** no supervisado
- **F-test:** supervisado.
- **Información mutua:** supervisado.



viu

Universidad
Internacional
de Valencia

universidadviu.com

De:
 Planeta Formación y Universidades