

APRENDIZAJE AUTOMATICO PROFUNDO (DEEP LEARNING)



□ **Profesora : Dra. Laura Lanzarini**

- Temas: Redes Neuronales y Técnicas de Optimización
- Aplicaciones en Minería de Datos y Procesamiento de Señales.

□ **JTP : Esp. César Estrebou**

- Temas: Redes Neuronales Profundas
- Desarrollo de aplicaciones de Machine Learning para Sistemas Embebidos.

□ **Ayudante : Ing. Marcos Saavedra**

- Becario doctoral CONICET

Bibliografía

- **Deep Learning with Python, 2nd edition.**

François Chollet.

Manning Publications Co. 2021

- **Neural Networks and Deep Learning**

Michael A. Nilsen

Determination Press. 2015

Semana	Fecha	Teoría	Práctica	Cuestionarios
1	17-ago	Introducción al aprendizaje automático y aprendizaje profundo.	P1.a) Revisión de las librerías básicas de Python	C1 - Preprocesamiento y visualización (habilitado del 26/8 al 8/9)
2	24-ago	Visualización y preprocesamiento.	P1.b) Preparación de datos para ML con Python	
3	31-ago	Redes Neuronales. Introducción. El perceptrón Matriz de confusión. Precisión y recall.	P2) Resolución de problemas linealmente separables	C2 - Perceptrón (habilitado del 9/9 al 22/9)
4	07-sep	Aprendizaje supervisado. Combinador lineal. Descenso del gradiente. Regresión polinomial.	P3.a) Minimización de funciones por gradiente. Resolución de problemas de regresión lineal y polinomial.	
5	14-sep	Neurona no lineal. Regresión Logística.	P3.b) Resolución de problemas de clasificación binaria.	C3 - Regresión y Clasificación binaria (habilitado del 23/9 al 6/10)
6	21-sep	Red Neuronal multiperceptrón. Algoritmo Backpropagation. Funciones de activación.	P4.a) MLP aplicado a la resolución de problemas concretos	
7	28-sep	Validación de modelos predictivos. Matriz de confusión. F-measure. Clasificación binaria. Curva ROC. Comparación de modelos usando AUC.	P4.b) validación de los modelos generados	C4 - MLP (habilitado del 7/10 al 20/10)
8	05-oct	Redes Neuronales Profundas. Lenguajes tensoriales. Visualización de la red. Tipos de capas. Funciones de pérdida. Backpropagation.	P5) Lenguajes tensoriales y tipos de capas	
9	12-oct	Redes convolucionales	P6.a) RN Convolucionales	C5- Redes convolucionales y redes recurrentes - OPCIONAL (habilitado del 21/10 al 11/11)
10	19-oct	Redes LSTM. Predicción de series temporales.	P6.b) resoluciones de problemas concretos	
11	26-oct	Consultas para la 1ra. Fecha		
12	02-nov	1ra. Fecha de Examen		

Reglamento

□ **ACTIVIDADES**

- ▣ Responder cuestionarios.
- ▣ Examen escrito al final del curso.

□ **NOTA FINAL** del curso

Promedio de

- ▣ Nota promedio de los cuestionarios.
- ▣ Nota del examen final

Reglamento

□ **ACTIVIDADES**

- ▣ Responder cuestionarios.
- ▣ Examen escrito al final del curso.

□ **NOTA FINAL** del curso

Promedio de

- ▣ Nota promedio de los cuestionarios.
- ▣ Nota del examen final

APROBACION DEL CURSO

□ **Promoción**

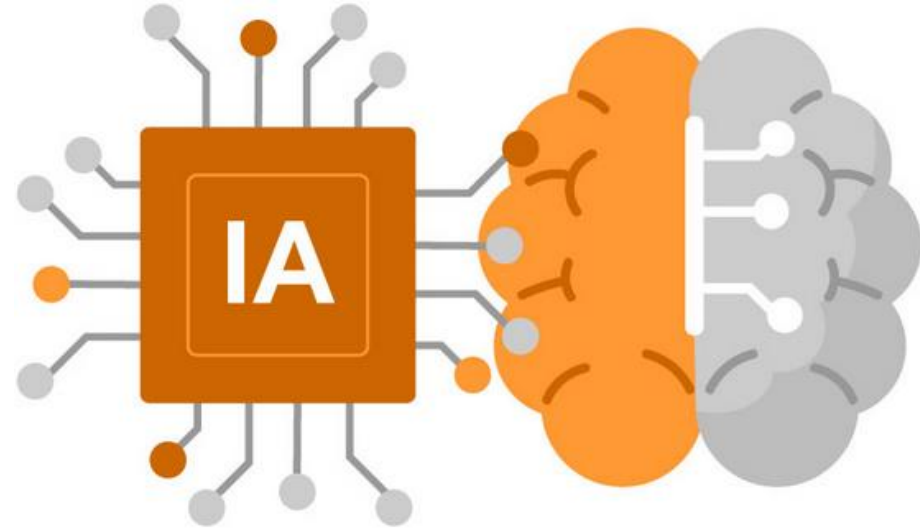
- ▣ 75% de los cuestionarios aprobados
- ▣ Nota examen escrito ≥ 6 puntos.
- ▣ NOTA FINAL ≥ 6 puntos.

□ **Cursada**

- ▣ 50% de los cuestionarios aprobados
- ▣ Nota examen escrito ≥ 4 puntos.
- ▣ NOTA FINAL ≥ 4 puntos

Inteligencia Artificial

- La **Inteligencia Artificial (IA)** es la inteligencia llevada a cabo por máquinas.
- **RAMAS**
 - ▣ **DEDUCTIVA** (lógica)
 - Sistemas expertos
 - ▣ **INDUCTIVA** (ejemplos)
 - Redes Neuronales
 - Técnicas de Optimización



Inteligencia Artificial

□ La **Inteligencia Artificial (IA)** es la inteligencia llevada a cabo por máquinas.

□ **RAMAS**

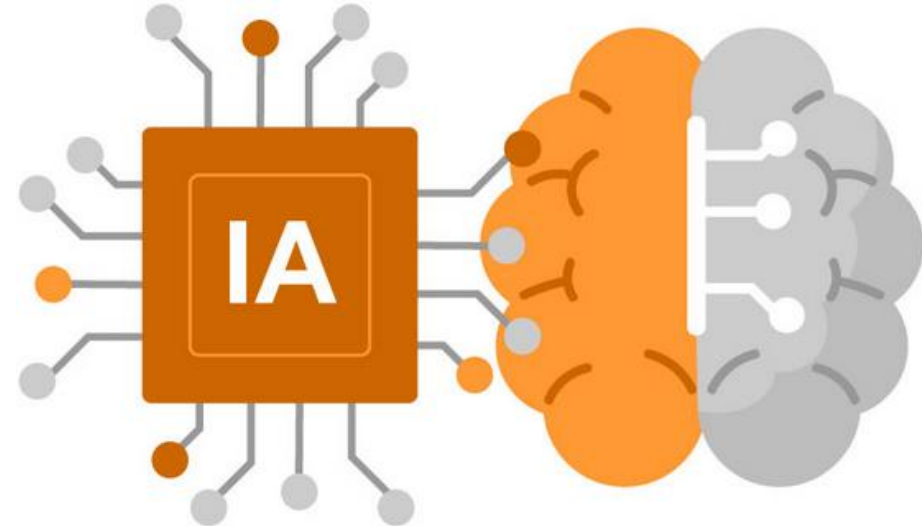
▣ **DEDUCTIVA** (lógica)

■ Sistemas expertos

▣ **INDUCTIVA** (ejemplos)

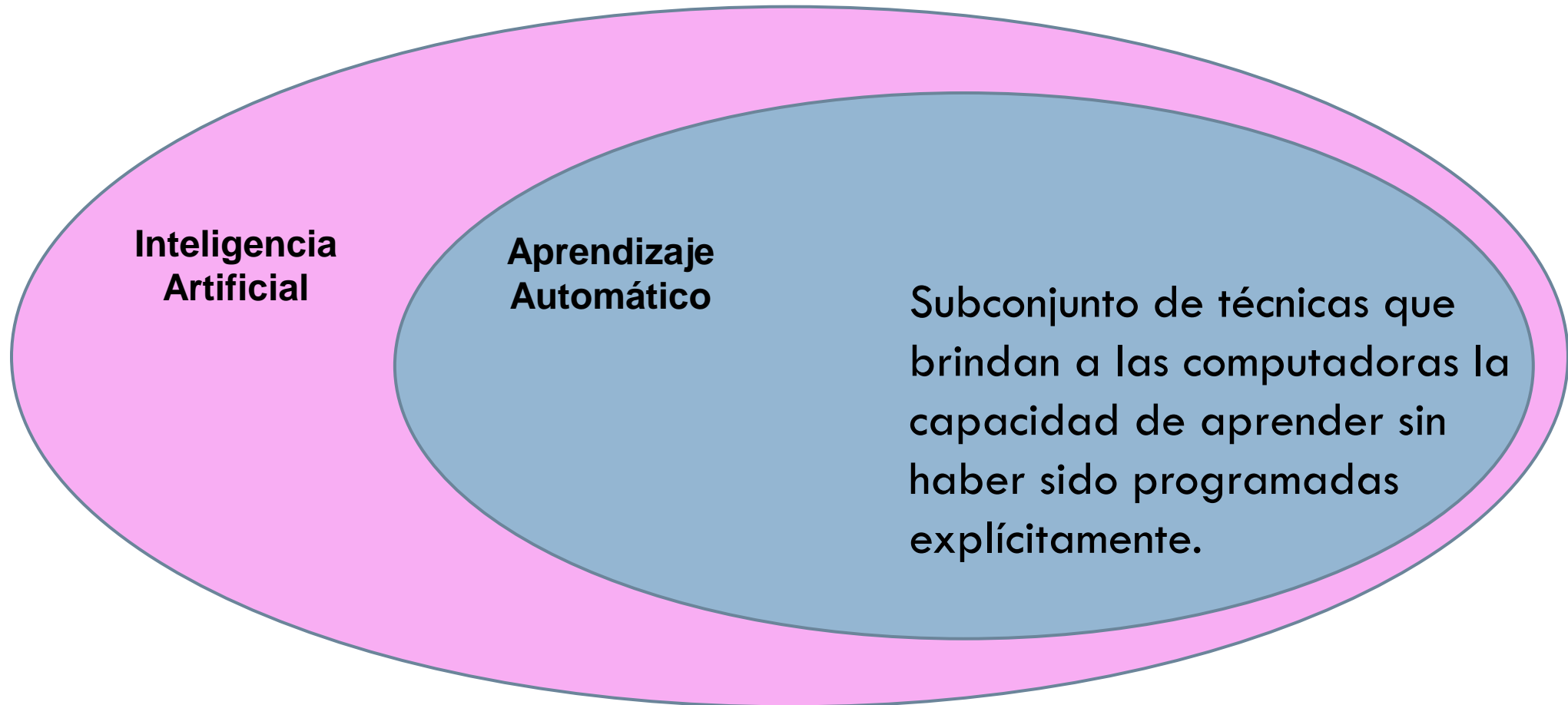
■ Redes Neuronales

■ Técnicas de Optimización

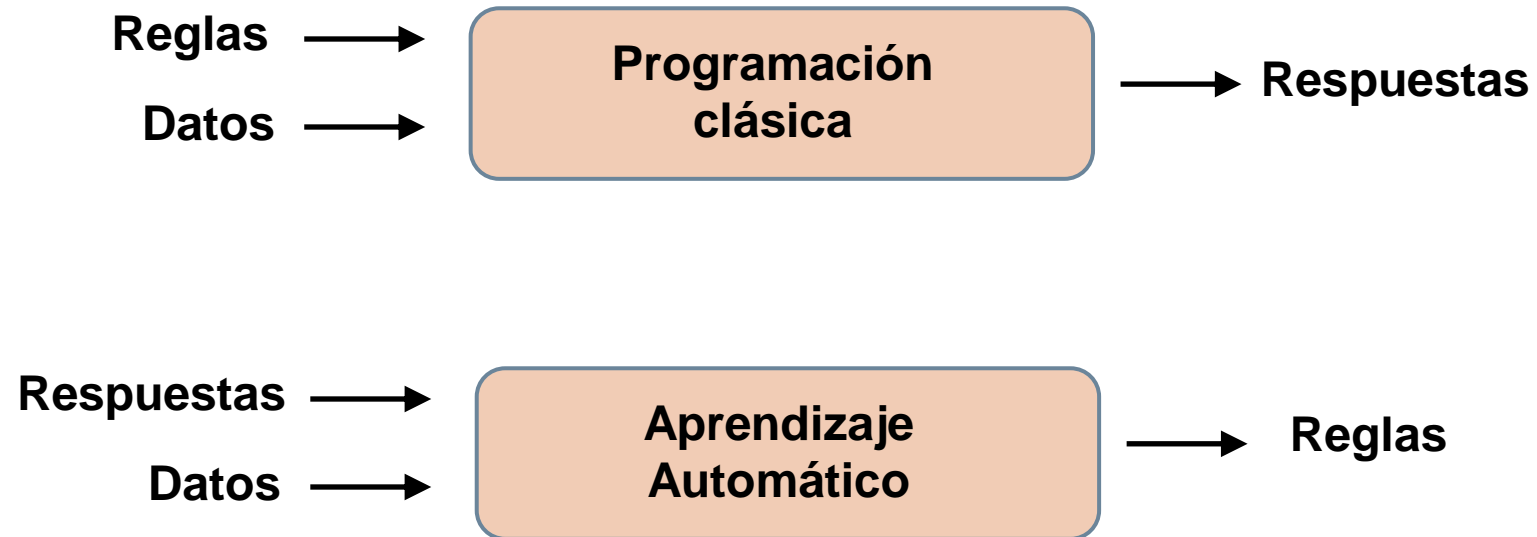


El **aprendizaje automático** pertenece a esta rama

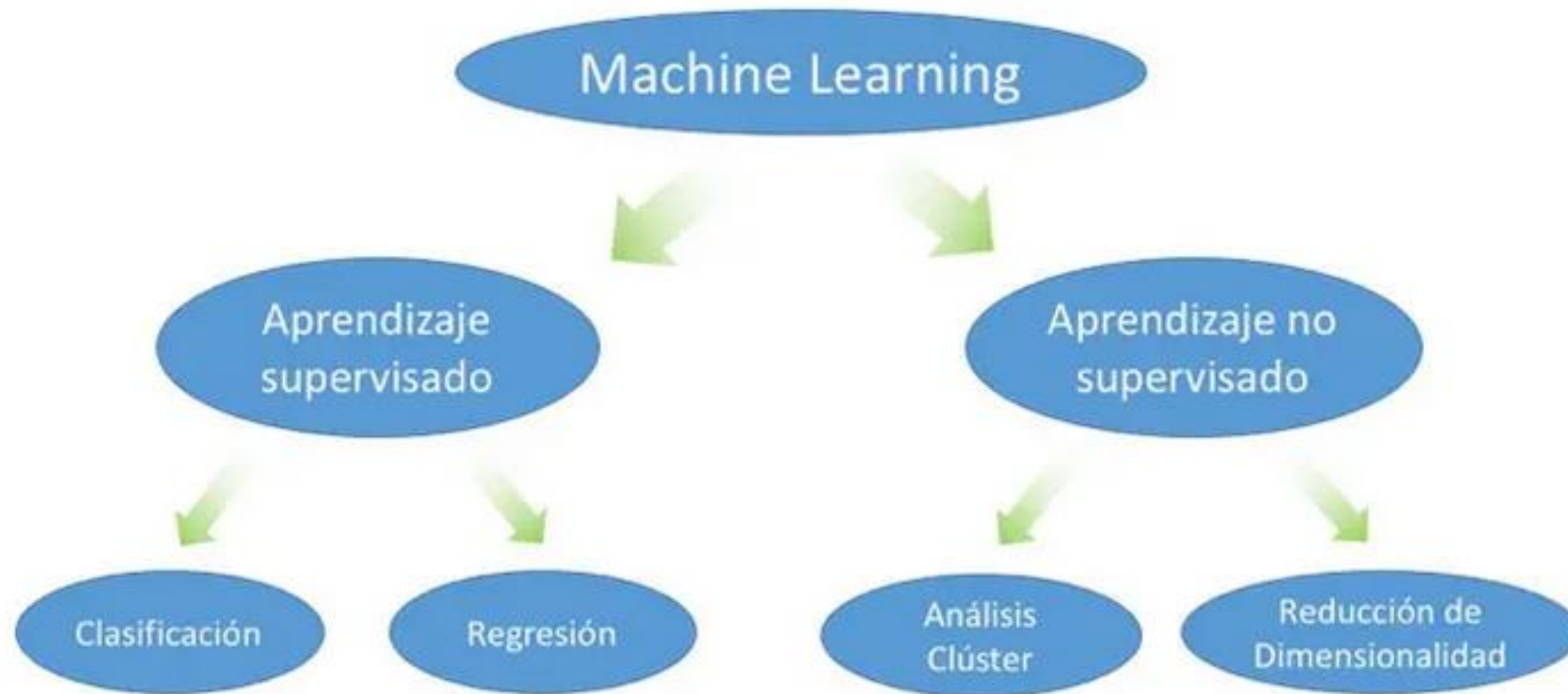
IA y Aprendizaje Automático



Programación clásica y Aprendizaje Automático



Tipos de aprendizaje



Aprendizaje supervisado

GATO



GATO



GATO



ARBOL



ARBOL



CUADERNO



CUADERNO



CUADERNO

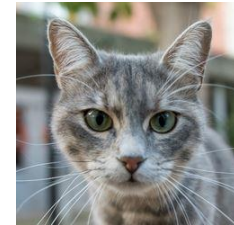


GATO



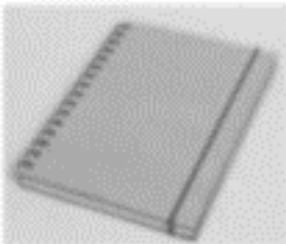
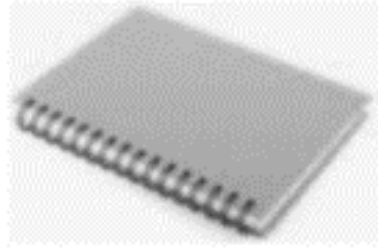
?

Aprendizaje no supervisado



AGRUPAMIENTO

Aprendizaje no supervisado



**Reducción de
características**

Aprendizaje supervisado

GATO



GATO



GATO



ARBOL



ARBOL



CUADERNO



CUADERNO



CUADERNO



**En este curso trabajaremos con
APRENDIZAJE SUPERVISADO**

GATO

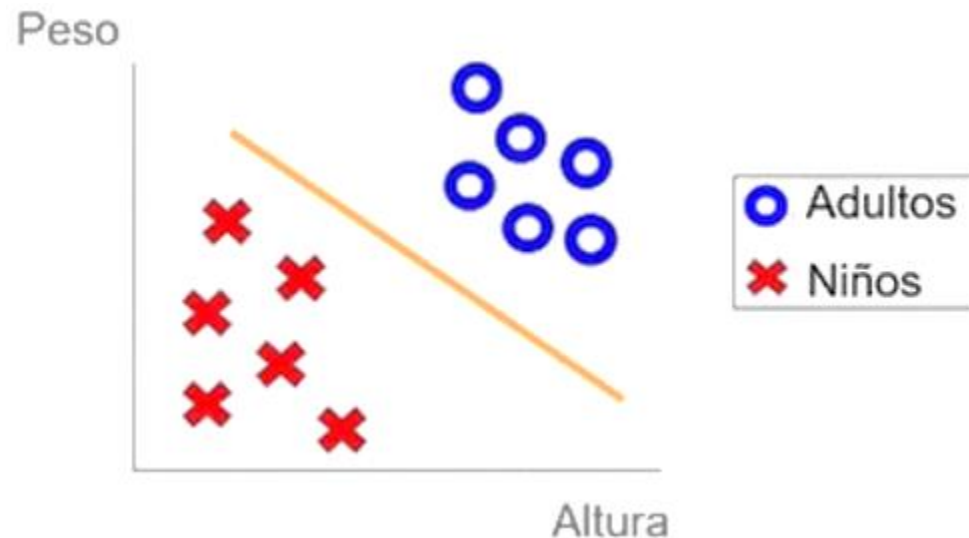


?

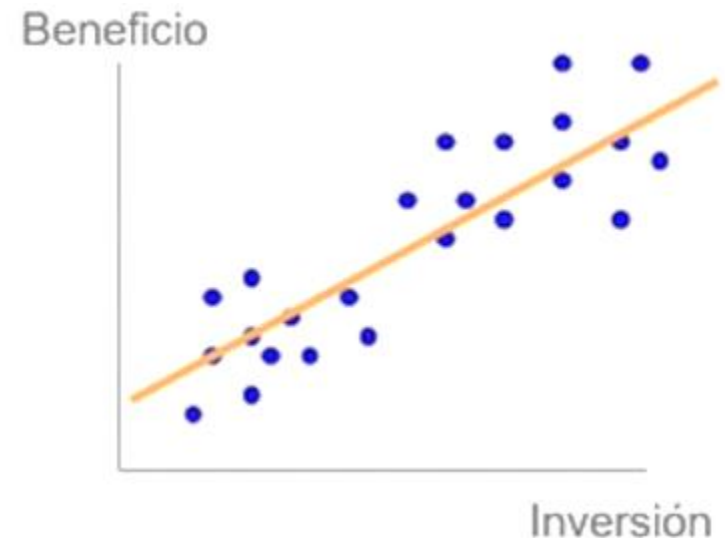
Aprendizaje Supervisado

- Según si la respuesta a predecir es **discreta** o **continua** se trata de un problema de **clasificación** o de **regresión** respectivamente.

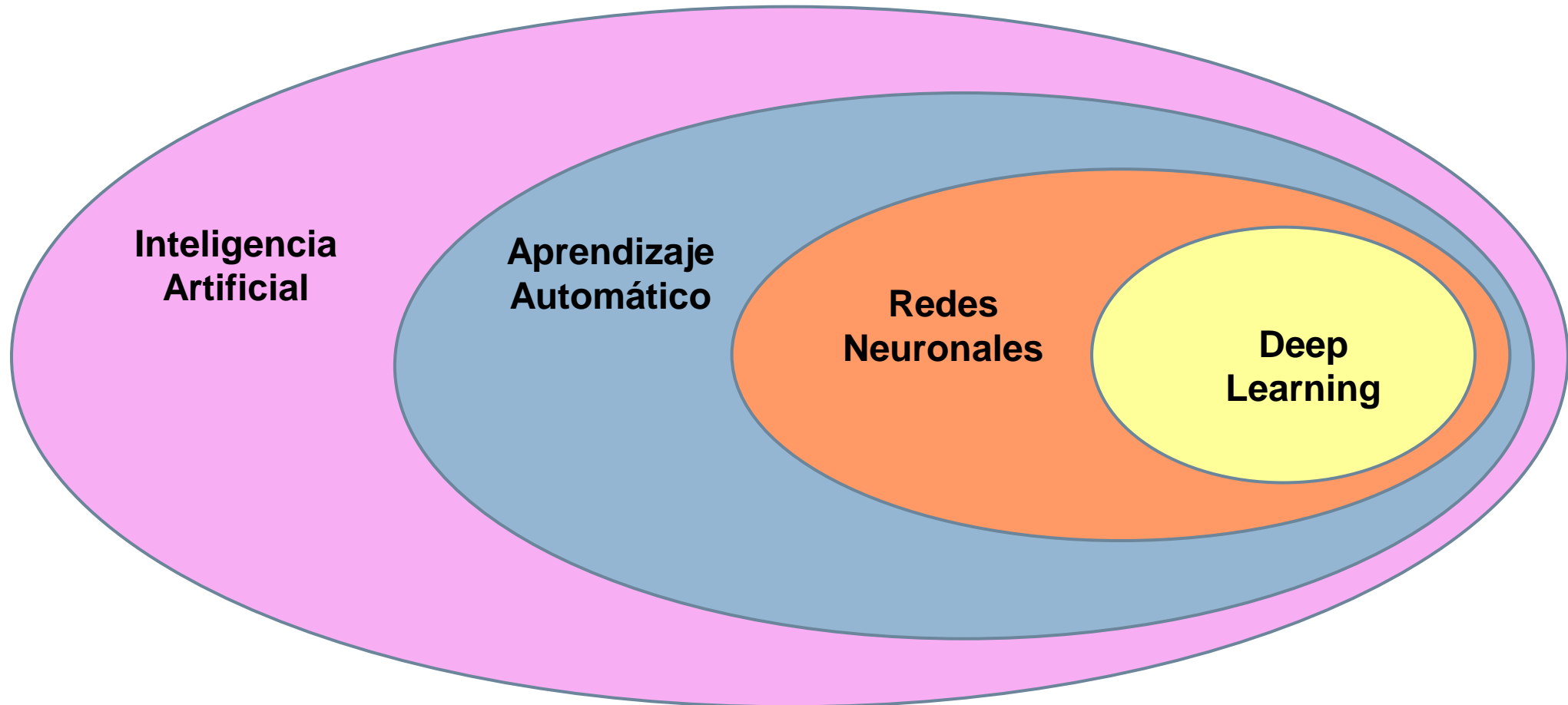
CLASIFICACION



REGRESION



Redes Neuronales y Aprendizaje Profundo



Tareas que pueden resolverse con RN

SUPERVISADO

- **Predicción** de un resultado futuro a partir de los datos disponibles.
 - ▣ Predecir el nivel de seguridad de un vehículo dadas sus características.
 - ▣ Determinar si un mail recibido es spam o no.
 - ▣ Dada la historia clínica de un paciente, predecir la probabilidad de contraer cierta enfermedad.

NO SUPERVISADO

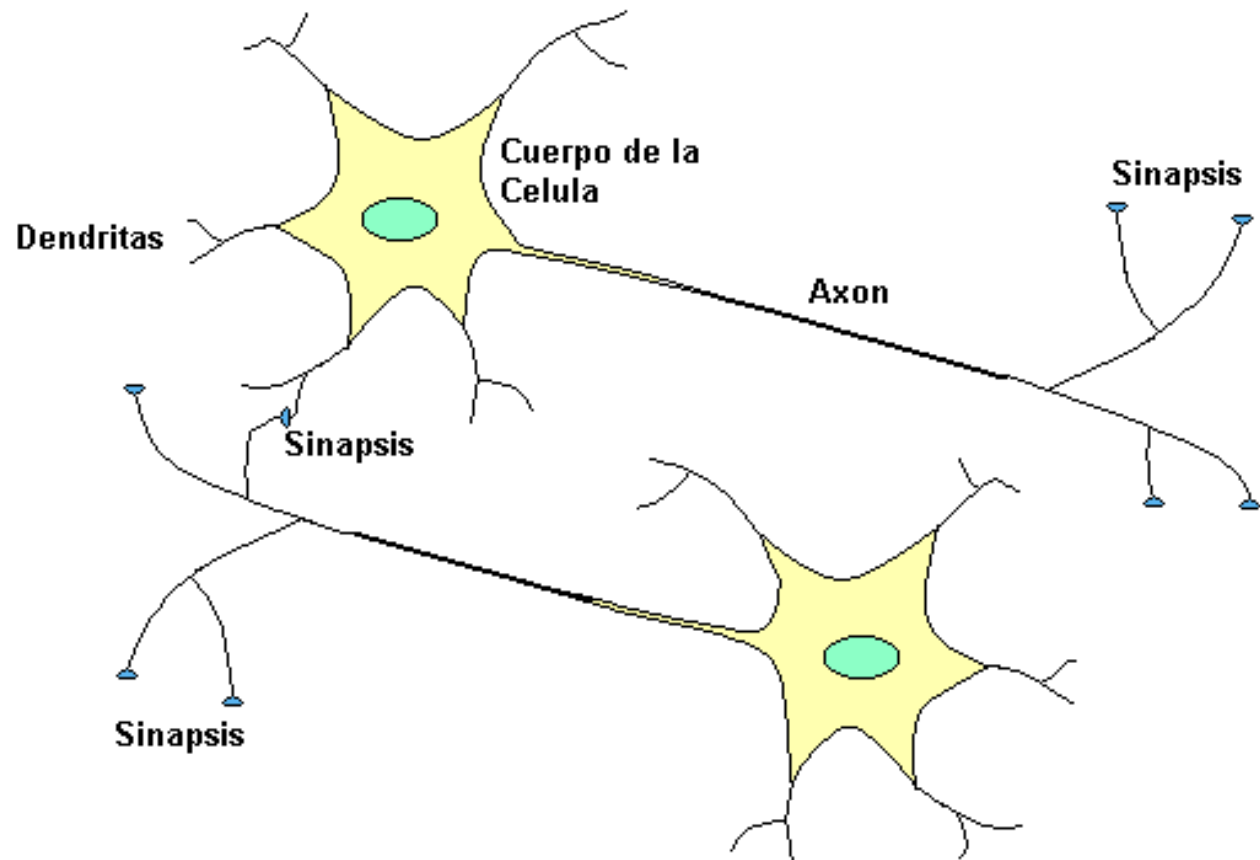
- **Segmentación** de los datos en subgrupos con características similares
 - ▣ Agrupar clientes para determinar perfiles que ayuden a direccionar campañas de marketing.
 - ▣ Caracterizar transacciones comerciales y detectar situaciones anómalas.

Redes Neuronales

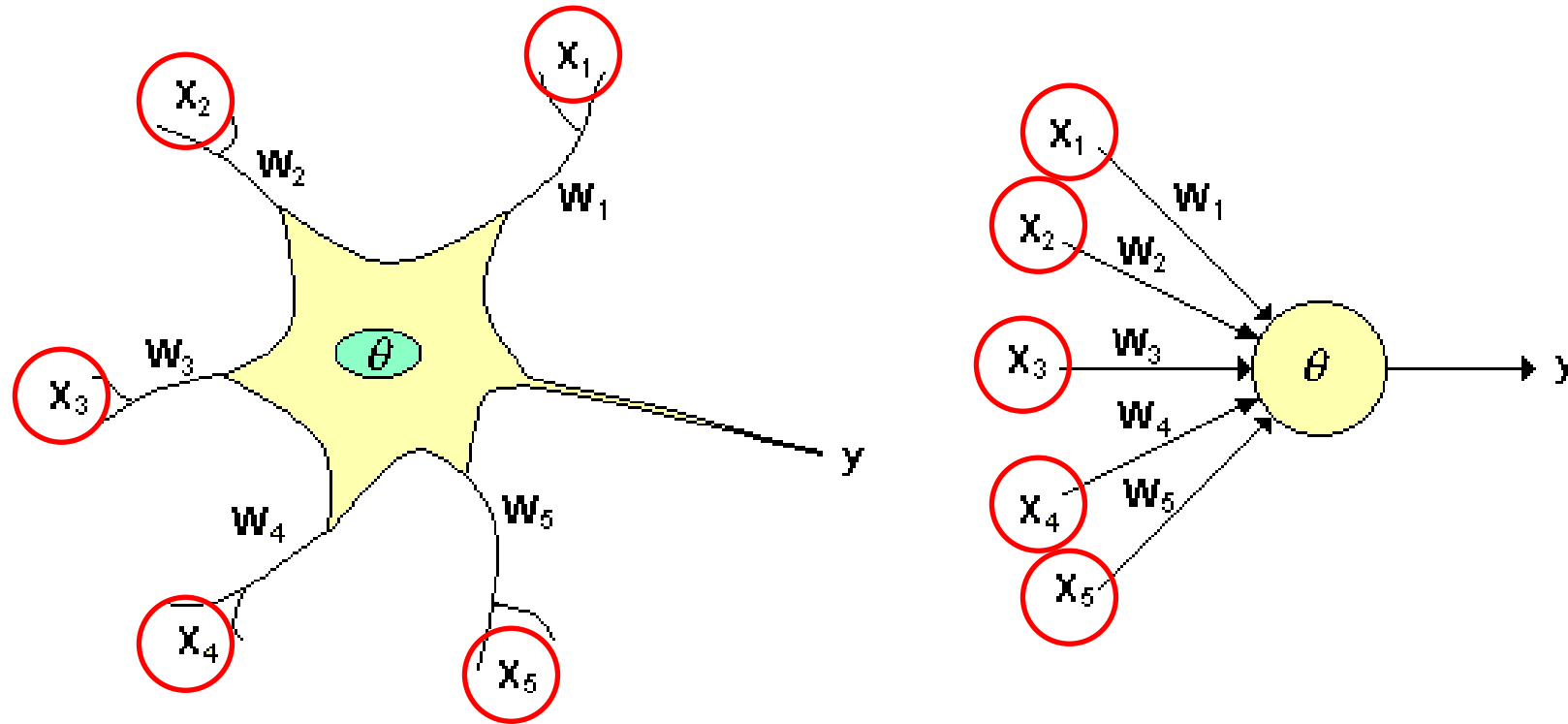
- El cerebro humano
 - ▣ Procesa información imprecisa rápidamente.
 - ▣ Aprende sin instrucciones explícitas.
 - ▣ Crea representaciones internas que permiten estas habilidades.
- Las Redes Neuronales Artificiales o simplemente **Redes Neuronales**, buscan emular el comportamiento del cerebro humano.

Neurona biológica

- El cerebro consta de un gran número de elementos (aprox. 10^{11}) altamente interconectados (aprox. 10^4 conexiones por elemento), llamados **neuronas**.

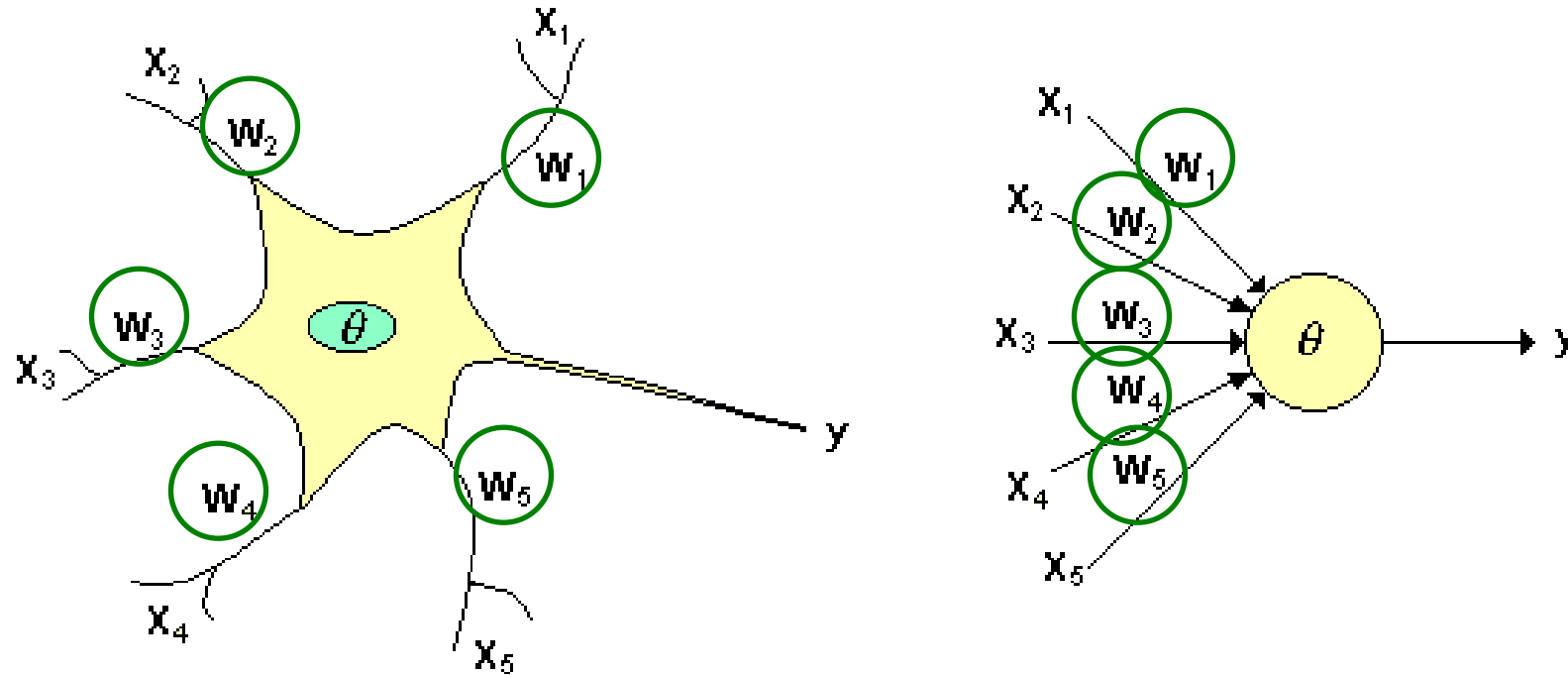


Similitudes entre una neurona biológica y una artificial



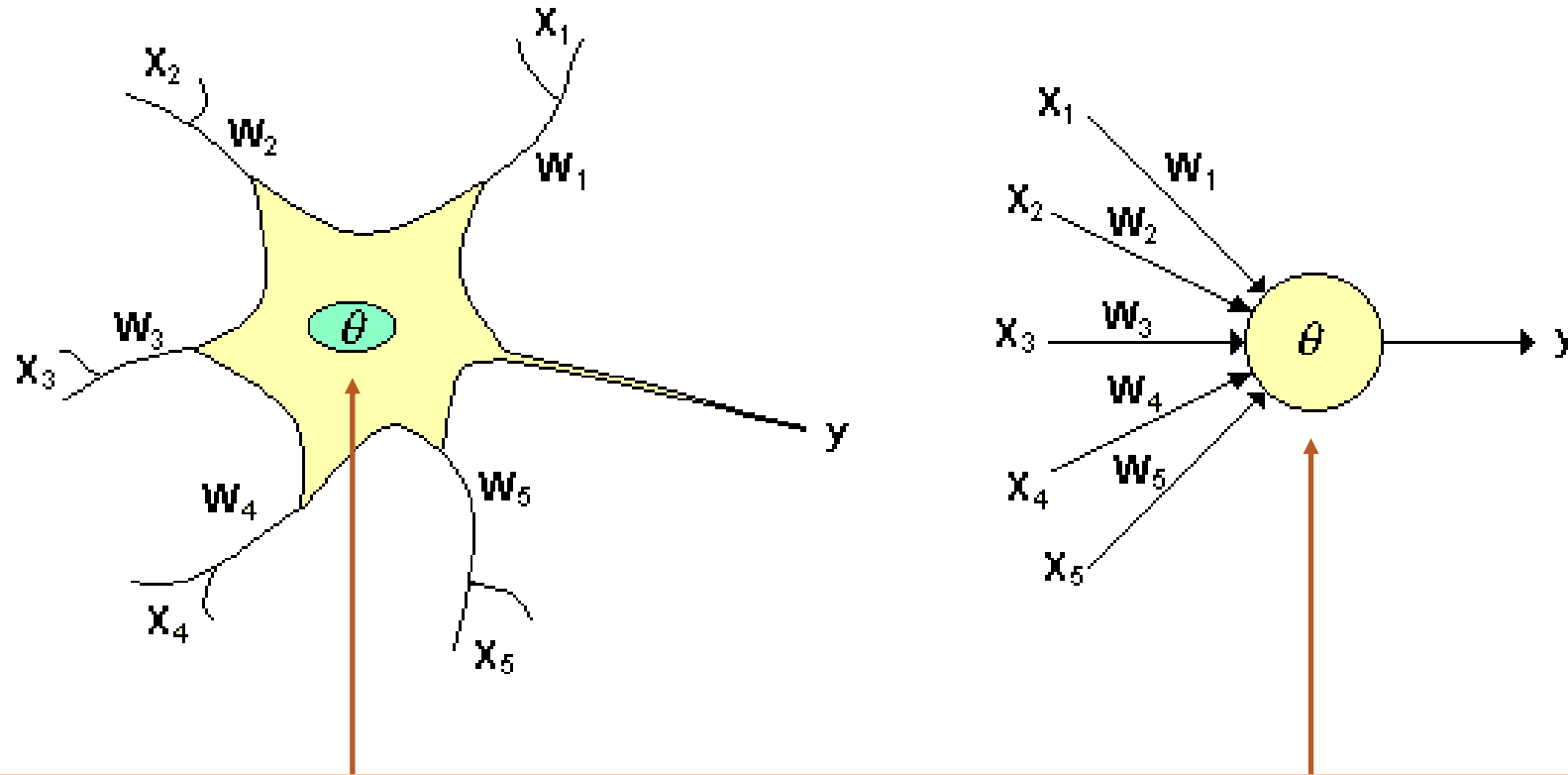
Las entradas X_i representan las señales que provienen de otras neuronas y que son capturadas por las dendritas

Similitudes entre una neurona biológica y una artificial



Los pesos W_i son la intensidad de la sinápsis que conecta dos neuronas; tanto X_i como W_i son valores reales.

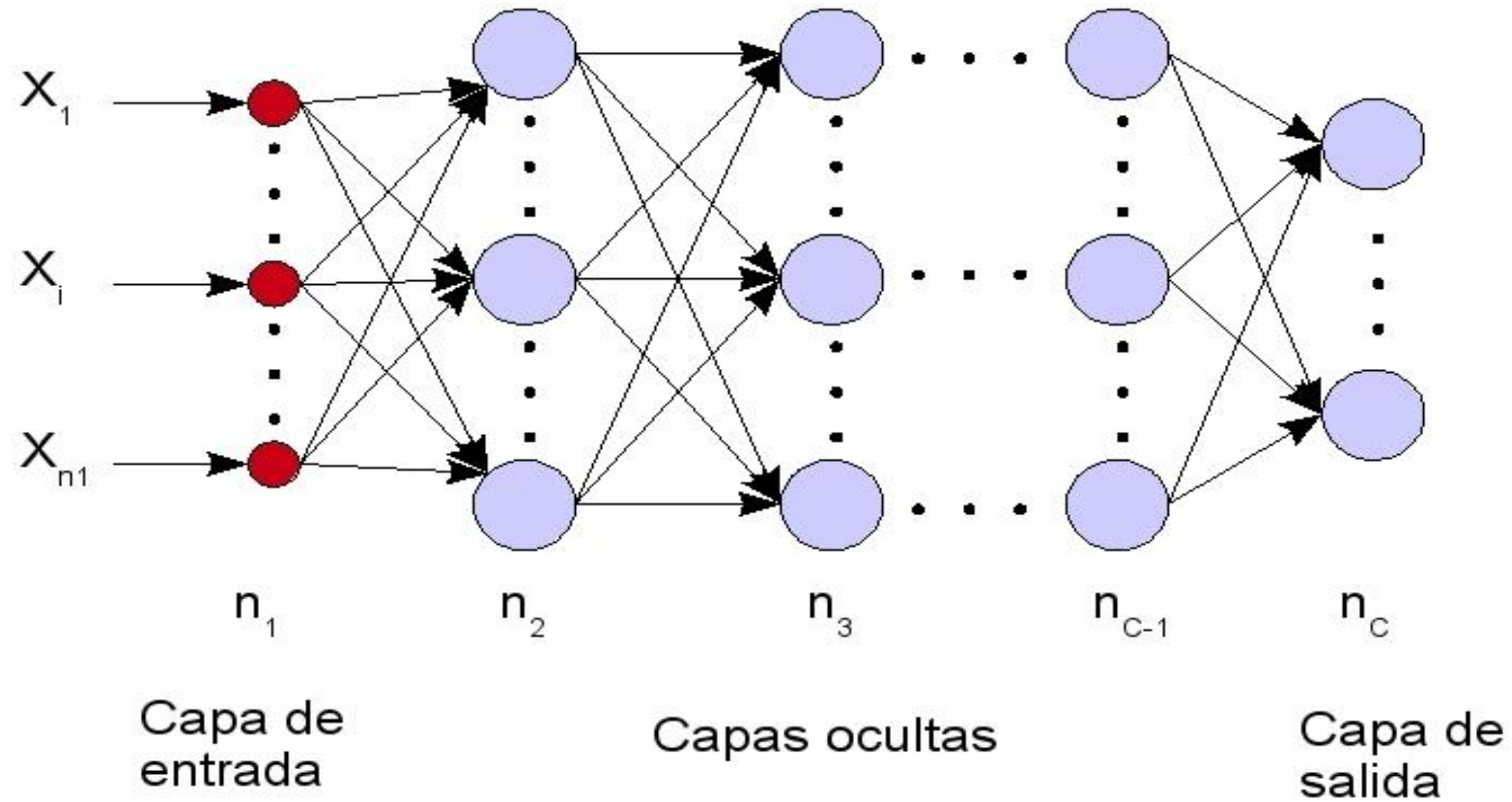
Similitudes entre una neurona biológica y una artificial



θ es la función umbral que la neurona debe sobrepasar para activarse; este proceso ocurre biológicamente en el cuerpo de la célula.

Red Neuronal Artificial

25



Orígenes de las RN

1943

Teoría de la Redes Neuronales

- ▣ Walter Pitts junto a Bertran Russell y Warren McCulloch intentaron explicar el funcionamiento del cerebro humano, por medio de una red de células conectadas entre sí.
- ▣ Lo aplicaron a la implementación de operaciones lógicas.
- ▣ Partieron del menor suceso psíquico (estimado por ellos): el impulso todo/nada, generado por una célula nerviosa.

Orígenes de las RN

1949

Conductividad de las sinapsis de las RN

- El fisiólogo Donald O. Hebb (de la McGill University) expuso que una percepción o un concepto se representa en el cerebro por un conjunto de neuronas activas simultáneamente.
- Afirmó que la memoria se localiza en las conexiones entre las neuronas (sinapsis).
- La regla de aprendizaje de Hebb presenta de manera intuitiva el modo en que las neuronas memorizan información. Esta regla indica que las conexiones entre dos neuronas se refuerzan si ambas son activadas.

Orígenes de las RN

1957

La primera Red Neuronal

- ▣ Frank Rosenblatt presentó el **Perceptron**, una red neuronal con aprendizaje supervisado cuya regla de aprendizaje era una modificación de la propuesta por Hebb.
- ▣ El principal aporte del Perceptron es que la adaptación de las conexiones entre las neuronas se realiza teniendo en cuenta el error entre la salida que da la red y la salida que se desea.
- ▣ En la fase siguiente de operación, la red es capaz de responder adecuadamente cuando se le vuelven a presentar los ejemplos de entrada.

Orígenes de las RN

1959

- ▣ **Widrow** publica una teoría sobre la adaptación neuronal y unos modelos inspirados en esa teoría, el Adaline (Adaptative Linear Neuron) y el Madaline (Multiple Adaline).
- Estos modelos fueron usados en numerosas aplicaciones y permitieron usar, por primera vez, una red neuronal en un problema importante del mundo real: filtros adaptativos para eliminar ecos en las líneas telefónicas.

1962

- ▣ **Roseblatt** utilizó la regla Delta como estrategia de aprendizaje.

Orígenes de las RN

1969

- ▣ Minsky y Papert demostraron las grandes limitaciones de esta red.
- PROBLEMA: Una red del tipo Perceptron no es capaz de aprender todas las posibles combinaciones entre entradas y salidas

70's

- ▣ No era bien visto trabajar con RN.

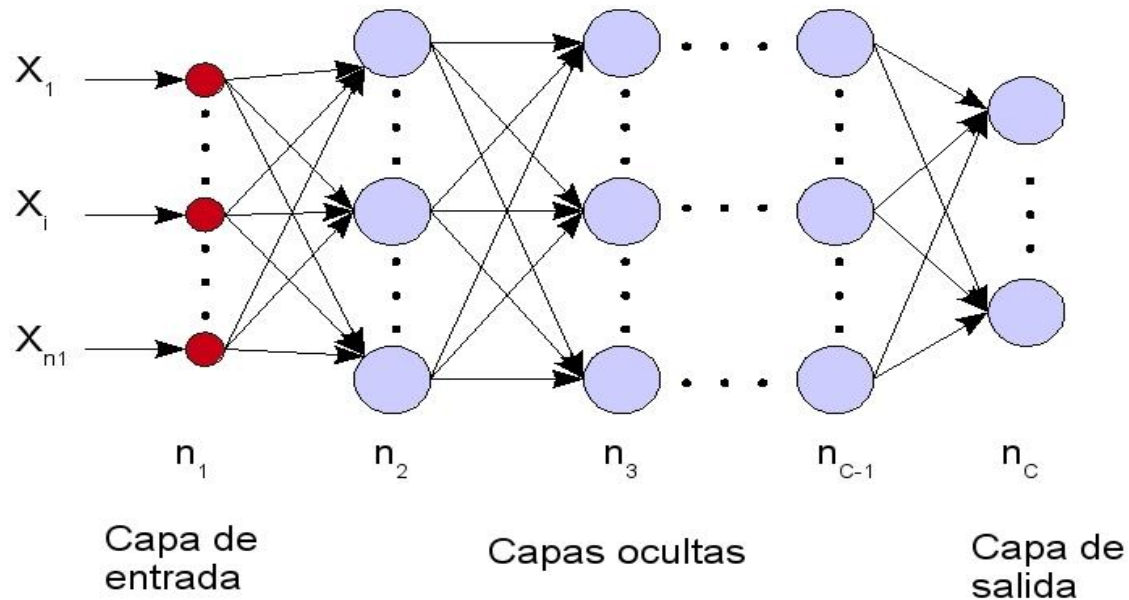
Orígenes de las RN

80's

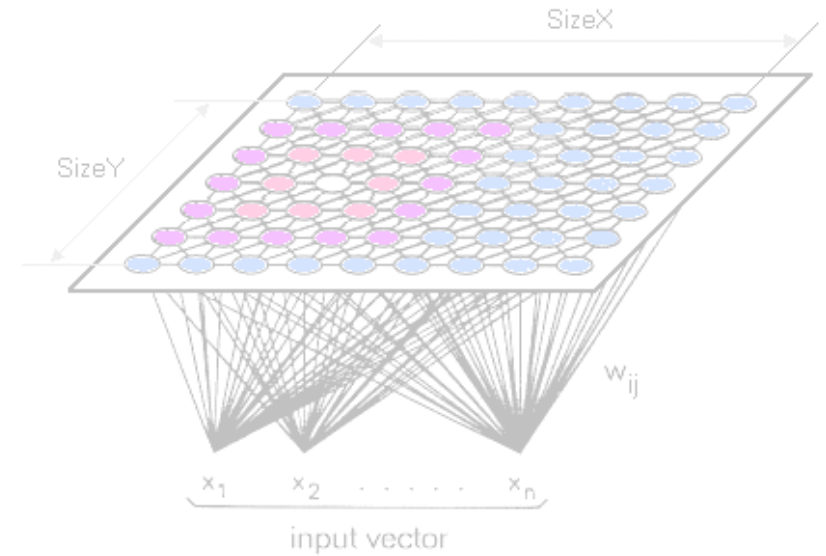
- ▣ Resurgen las Redes Neuronales con la aplicación del algoritmo Backpropagation descrito por Paul Werbos en 1974.
 - 1980 Proyecto DARPA (Defense Advanced Research Projects Agency).
 - 1983 Hopfield y los modelos BAM de Kosko reimpulsaron el tema.
 - 1987 IEEE International Conference on Neural Networks
 - 1988 Journal de la INNS (International Neural Networks Society)
 - 1990 IEEE Transaction on Neural Networks

Redes Neuronales. Arquitecturas clásicas

□ Predictiva



□ Descriptiva

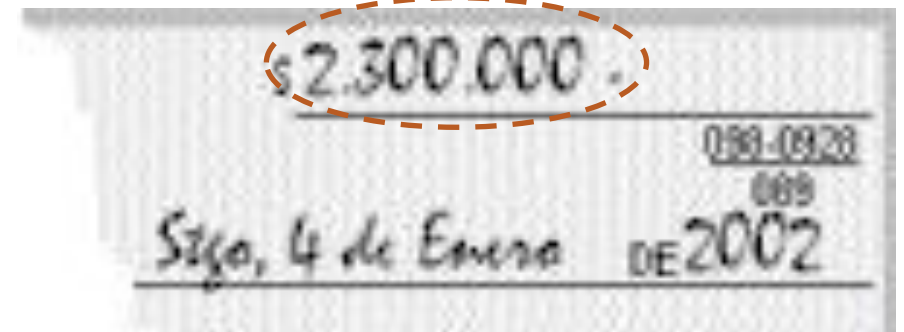
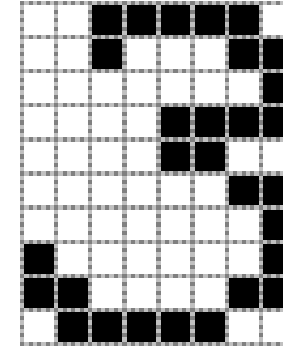
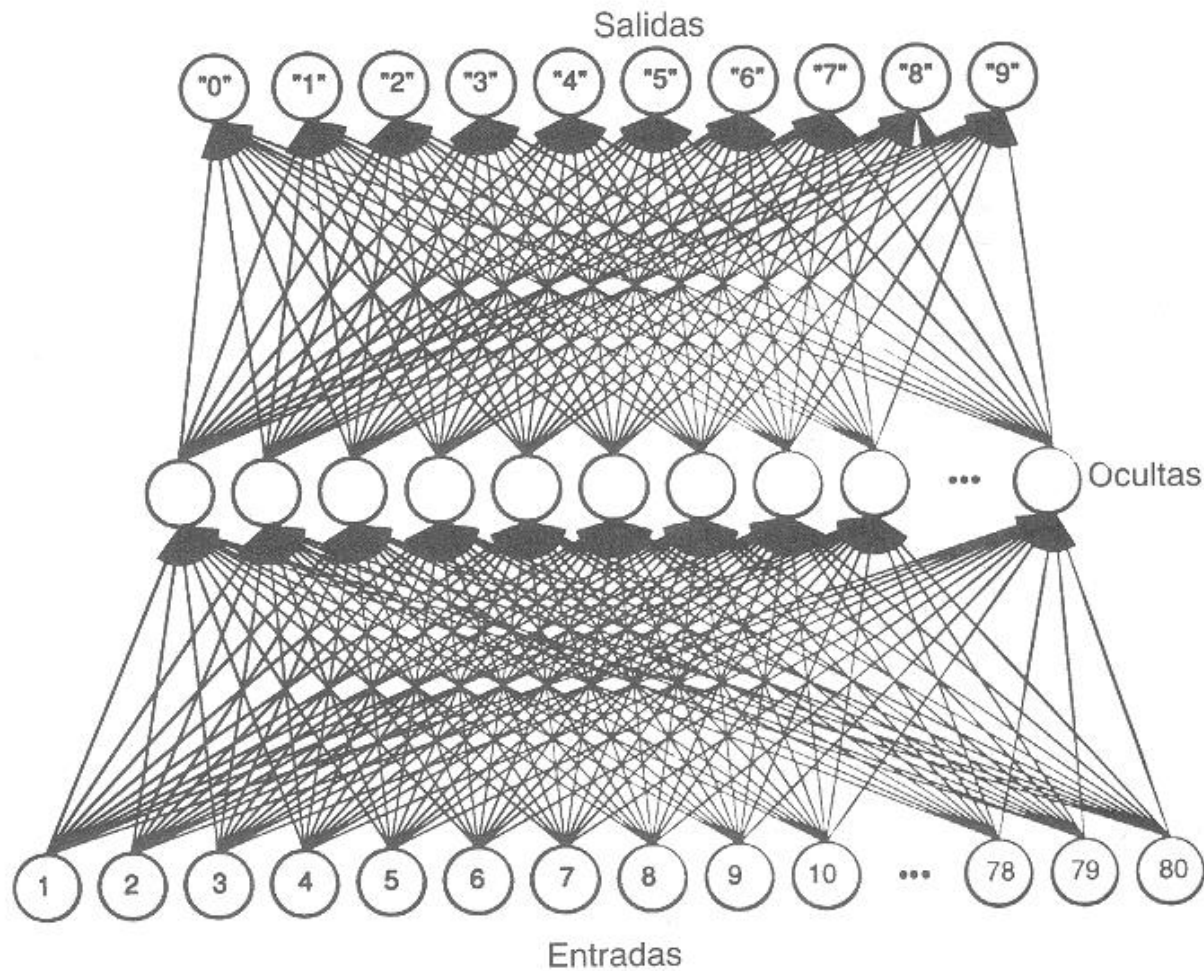


Ejemplo

- Se desea entrenar una red neuronal para que reconozca caracteres escritos a mano

5 3 4

Reconocimiento de dígitos manuscritos



Reconocimiento de dígitos manuscritos

35

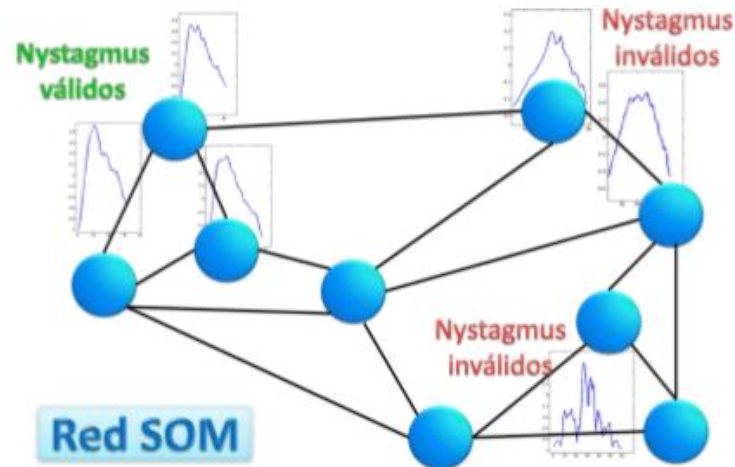
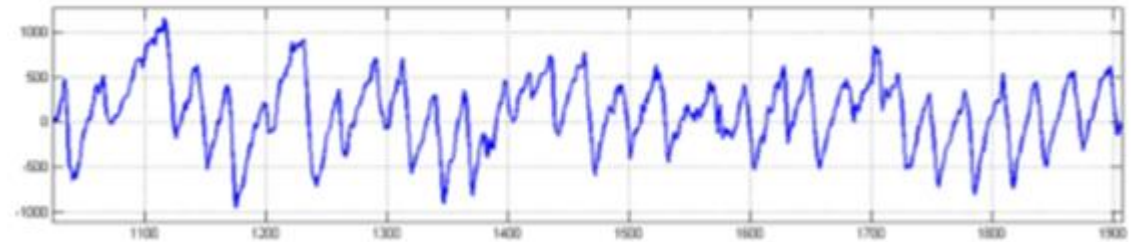
- Caracteres correctamente reconocidos

5 3 8 4

- Caracteres NO reconocidos

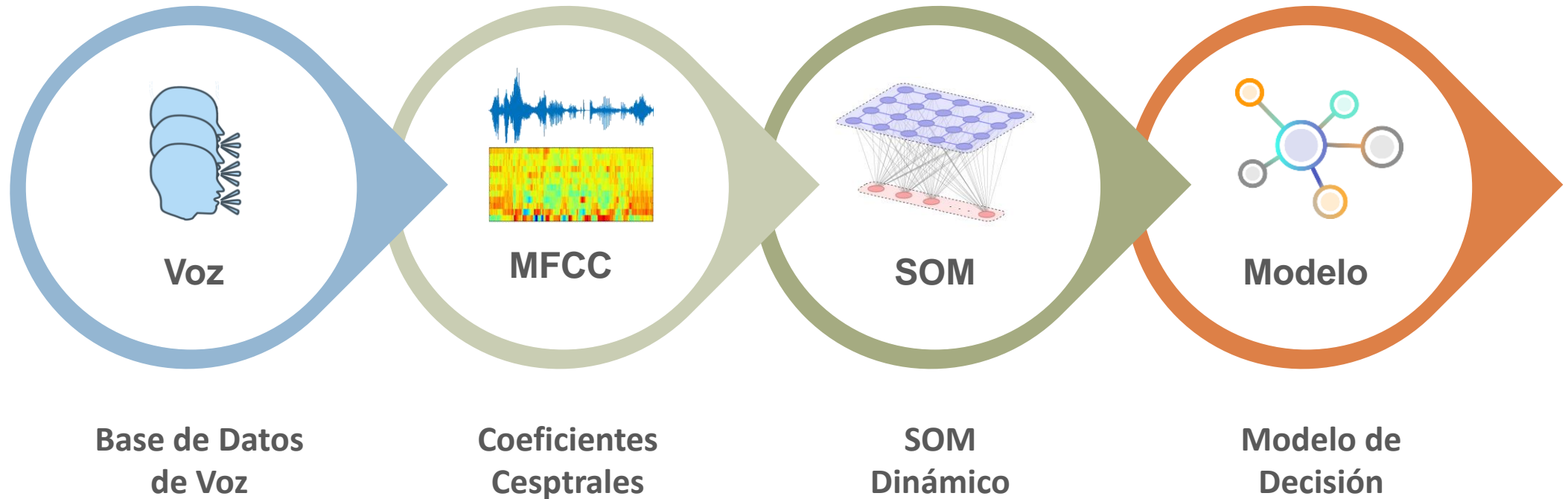
1 0 2

Diagnóstico de alteraciones del equilibrio



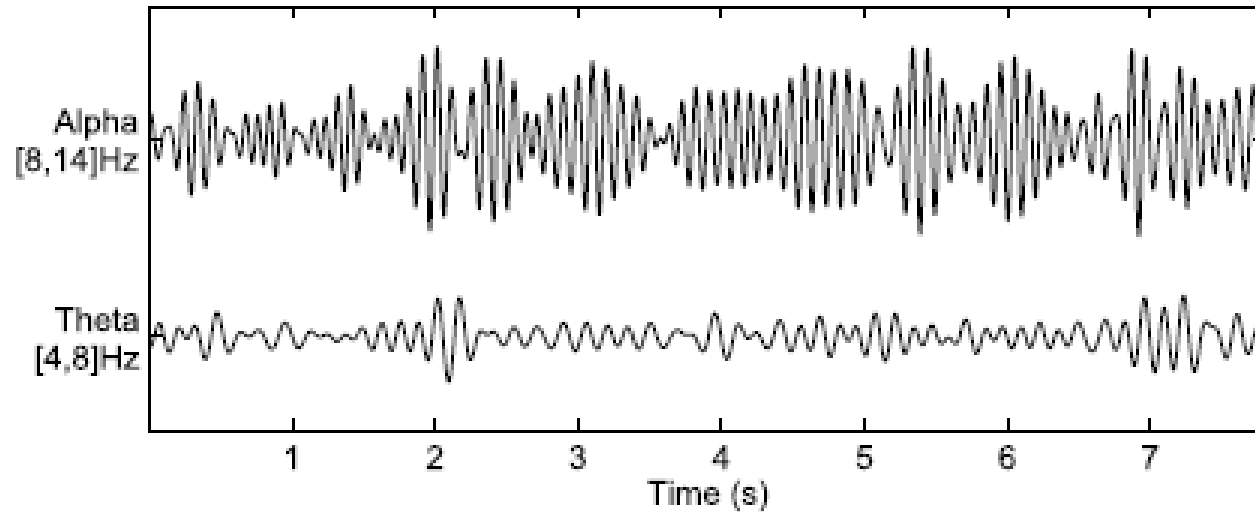
- Funcionando en consultorio.
- Realiza un prediagnóstico en forma automática.
- Registro de software

Reconocimiento de voz



- Voces de 30 locutores durante 20 seg. para entrenar.
- Cada segmento de audio se representa por una secuencia de coef. ceptrales
- Se usaron intervalos de 20 ms con superposición de 10 ms.
- La red usa un sistema de votación para responder.

Detección temprana de demencia



Luis Guerra et al.(2018). *The Electroencephalogram as a Biomarker Based on Signal Processing Using Nonlinear Techniques to Detect Dementia*. In: *Developments and Advances in Defense and Security. MICRADS 2018. Smart Innovation, Systems and Technologies*, vol 94. Springer.

https://doi.org/10.1007/978-3-319-78605-6_11

Sistemas inteligentes



Sistema Inteligente de Transporte – SITBus (billete electrónico) + SAO (control operativo) + SIU (información al usuario)



Análisis de imágenes

- Pinterest incorporó **VisualGraph**



Detector de personas



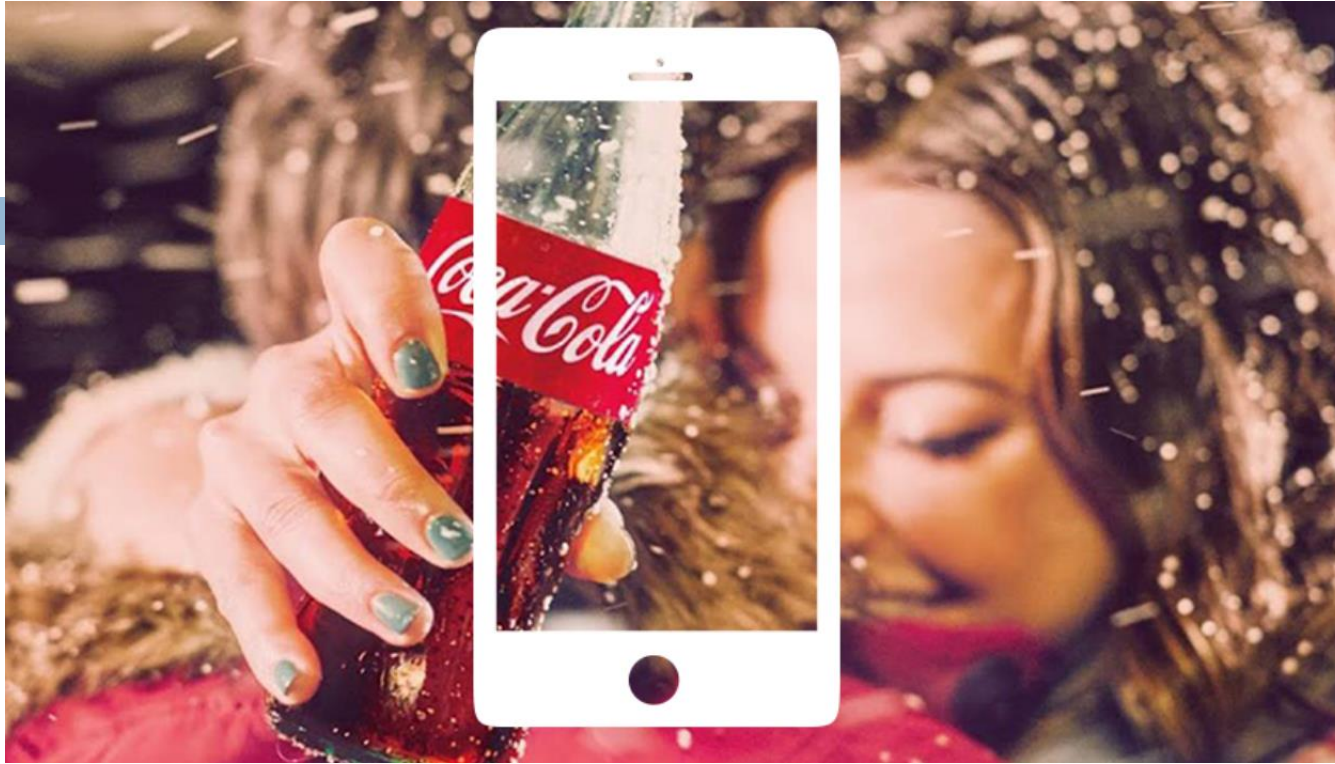
Detector de bolsos



Detector de faldas

<https://techcrunch.com>

- Empresa **Vicarious** : Inversores Mark Zuckerberg (Facebook), Elon Musk (cofundador de PayPal) buscan determinar las “relaciones de causa y efecto”.
- 2.300 millones de usuarios activos en Facebook generando muchos datos. (Fuente: Data Never Sleeps 2019)

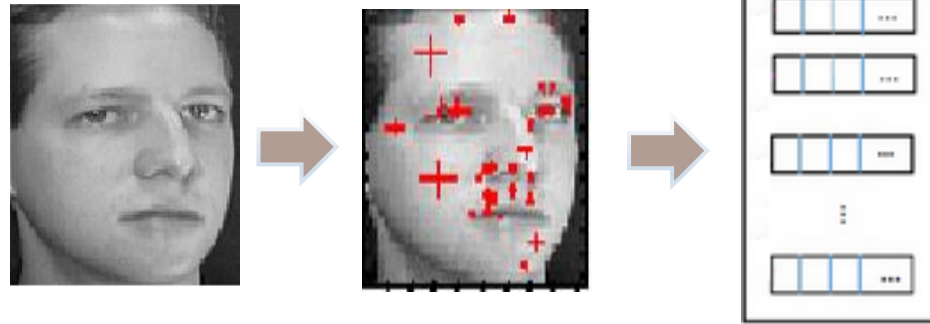


www.adweek.com

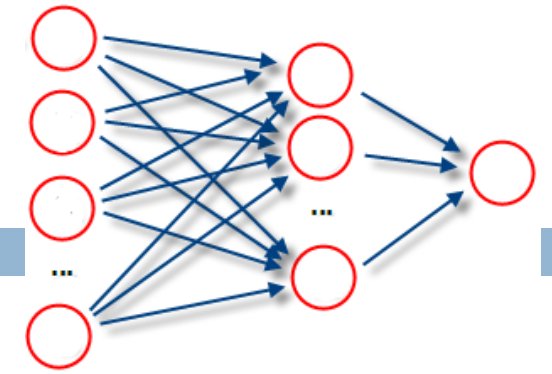
- Seguimiento de sus redes sociales para saber
 - ▣ quién está consumiendo sus bebidas
 - ▣ dónde están sus clientes
 - ▣ qué situaciones los incitan a hablar sobre su marca
- Identifica sus productos en fotografías y determina cuando enviar publicidad
- Ahora buscan usar bots para generar anuncios

Representación de los datos

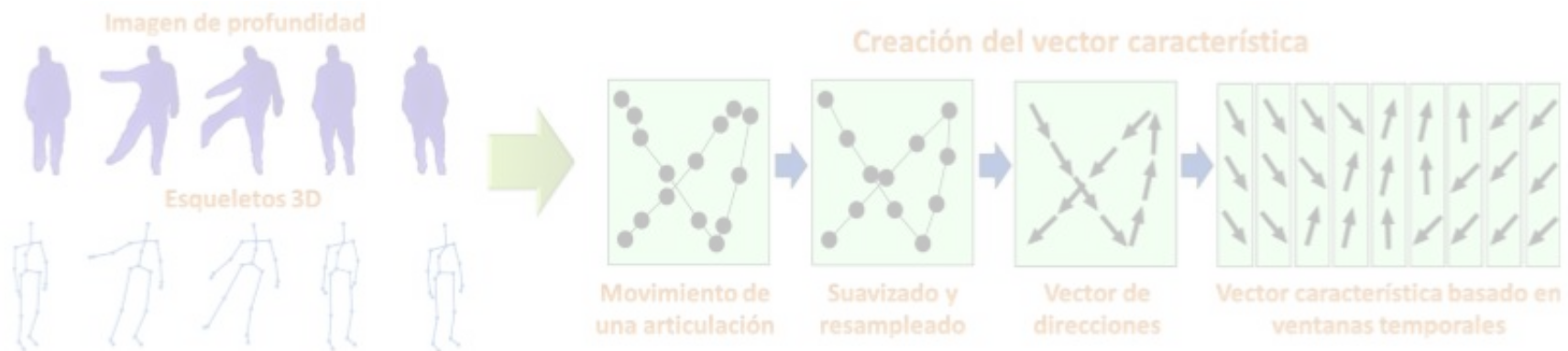
□ Caracterización de rostros



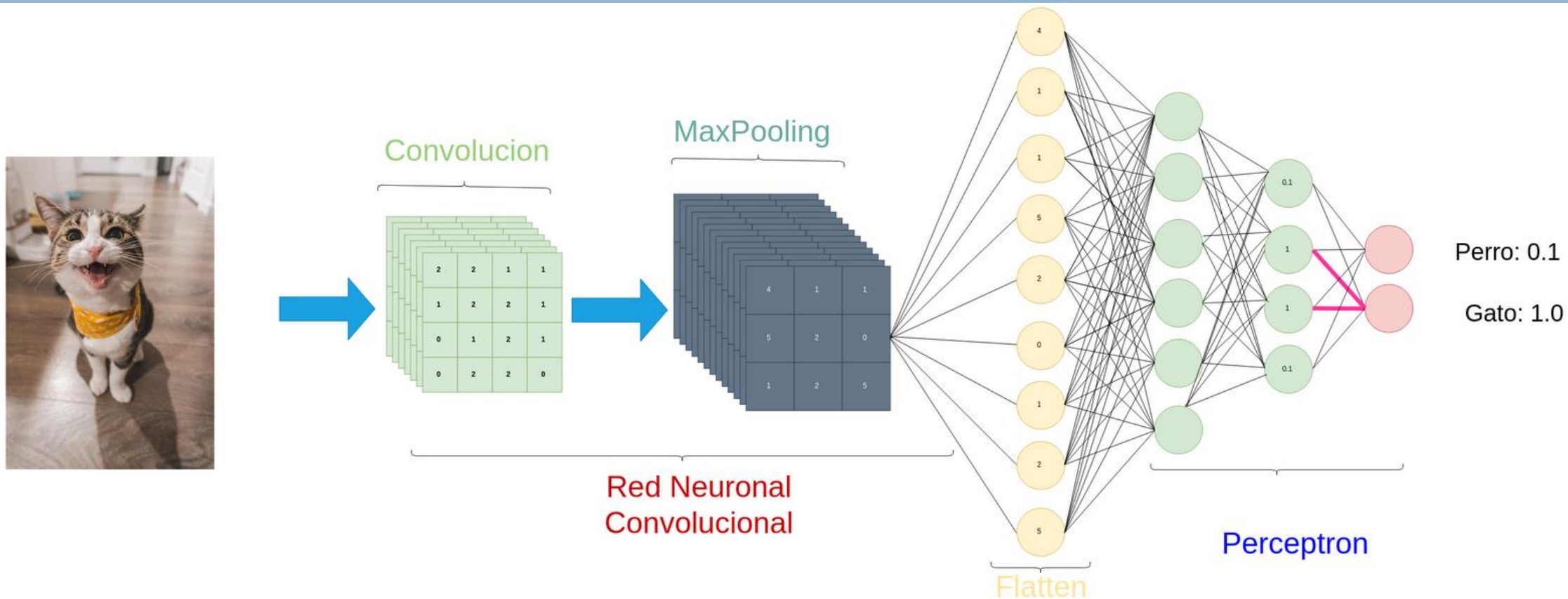
SIFT features - Lowe (2004)



□ Gestos Dinámicos



Redes Neuronales Convolucionales

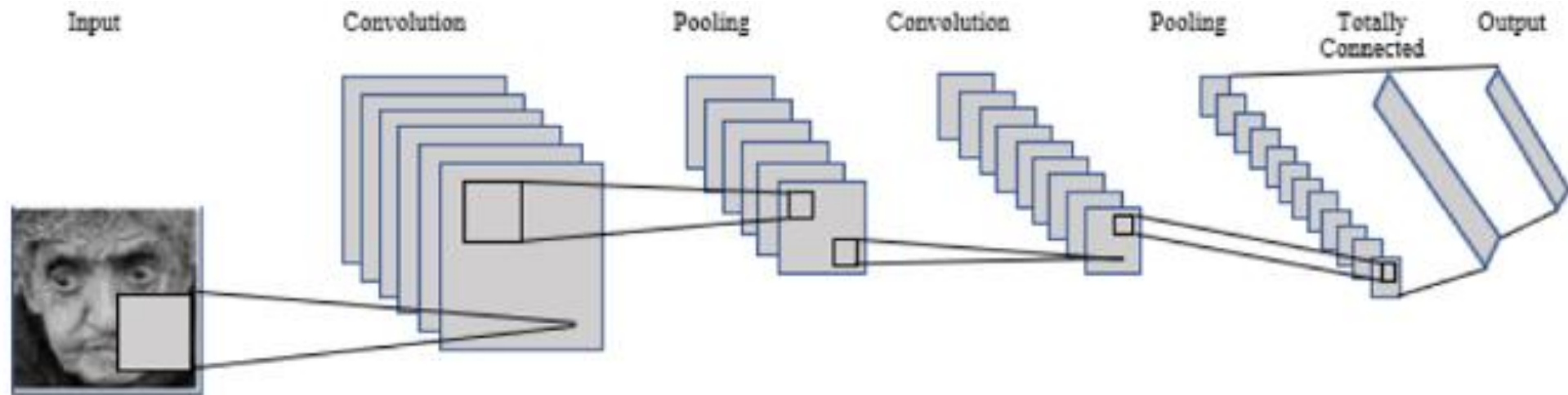


Reconocimiento de expresiones faciales



- BBDD Facial Expressions in the Wild (+ de 80 mil imágenes.
Alegría, sorpresa, tristeza, enojo, miedo y disgusto)
- Arquitecturas de CNNs : VGG, Inception o ResNet
- TensorFlow, Keras y PyTorch (Frameworks para Deep Learning)

Expresiones faciales en pacientes con Alzheimer



Castillo-Salazar D. et al. (2020) **Detection and Classification of Facial Features Through the Use of Convolutional Neural Networks (CNN) in Alzheimer Patients.** In: *Human Systems Engineering and Design II. IHSED 2019. Advances in Intelligent Systems and Computing*, vol 1026. Springer.

https://doi.org/10.1007/978-3-030-27928-8_94

Redes Neuronales que generan datos

- 2014 □ **Redes Generativas Adversarias (GAN)** generan nuevos datos en situaciones en que éstos son limitados.



Redes Neuronales que generan datos

- 2014 □ **Redes Generativas Adversarias (GAN)** generan nuevos datos en situaciones en que éstos son limitados.

<https://dl.acm.org/doi/10.5555/2969033.2969125>

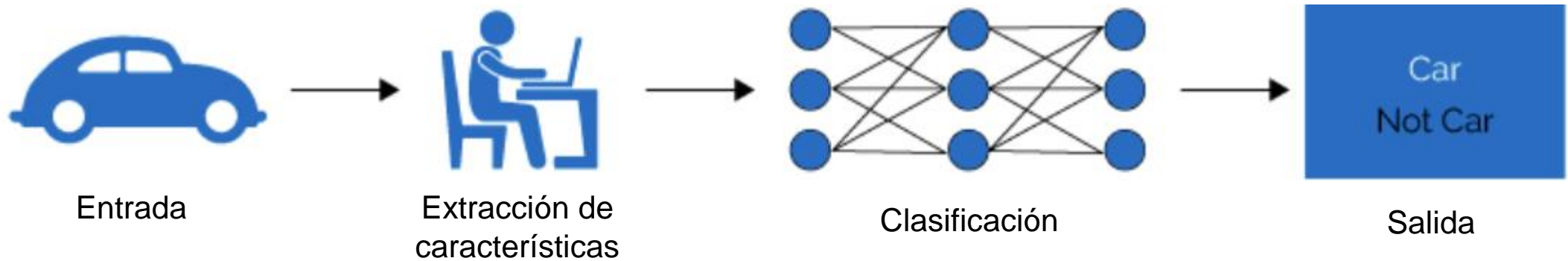
- 2019 □ **Autoencoders Variacionales (VAE)** tienen por objetivo reconstruir los datos de entrada.

- ▣ DeepMind demostró que los VAEs podían superar a las GAN en la generación de caras.

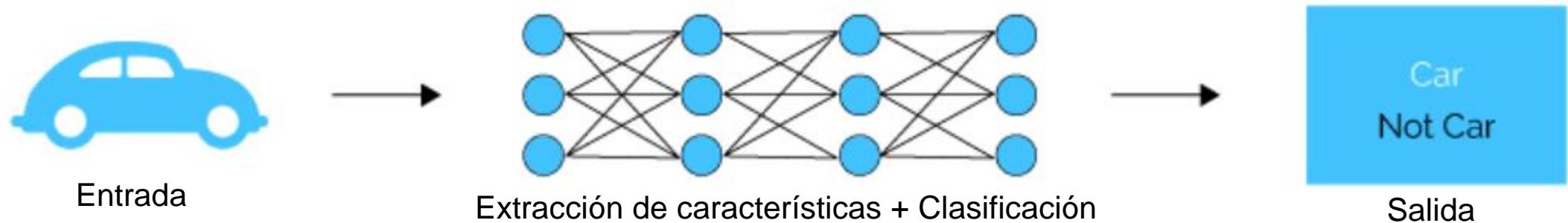
<https://arxiv.org/abs/1906.00446>

Contenido del curso

APRENDIZAJE AUTOMATICO



DEEP LEARNING



Redes Neuronales - Arquitecturas

PARTE I

- Perceptrón
- Combinador Lineal
- Neurona no lineal
- Multiperceptrón (aprendizaje backpropagation)

PARTE II

- Tensores y tipos de capas
- Funciones de pérdida
- Redes convolucionales
- Redes recurrentes

Ejemplo: Clasificación de flores de Iris

- Se dispone de información de 3 tipos de flores Iris



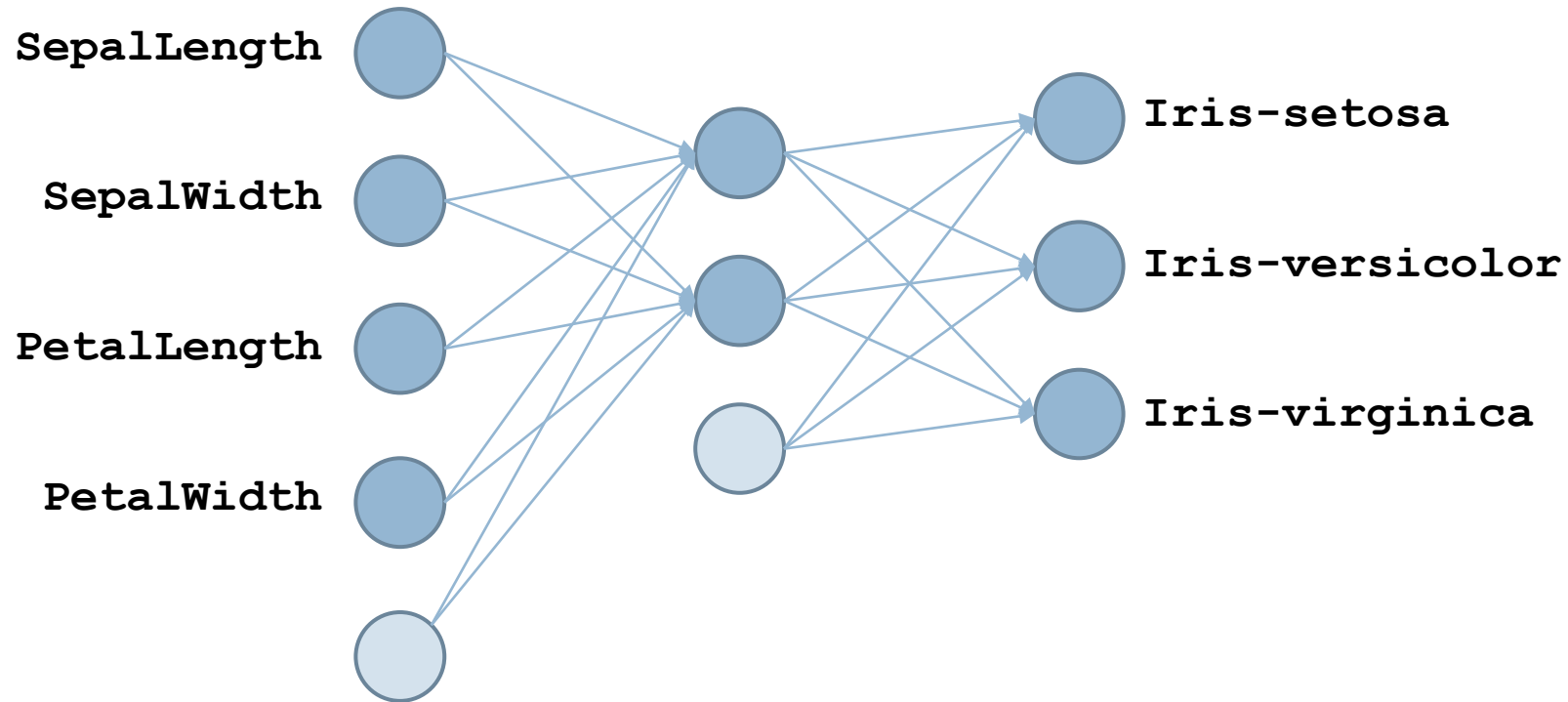
<https://archive.ics.uci.edu/ml/datasets/Iris>

Ejemplo: Clasificación de flores de Iris

Id	sepalength	sepalwidth	petallength	petalwidth	class
1	5,1	3,5	1,4	0,2	Iris-setosa
2	4,9	3,0	1,4	0,2	Iris-setosa
...
95	5,6	2,7	4,2	1,3	Iris-versicolor
96	5,7	3,0	4,2	1,2	Iris-versicolor
97	5,7	2,9	4,2	1,3	Iris-versicolor
...
149	6,2	3,4	5,4	2,3	Iris-virginica
150	5,9	3,0	5,1	1,8	Iris-virginica

<https://archive.ics.uci.edu/ml/datasets/Iris>

Ejemplo: Clasificación de flores de Iris



Ejemplo: Prescripción de lentes de contacto

- Se dispone de la siguiente información de pacientes atendidos previamente.
 - ▣ **EDAD** del paciente: joven, pre-presbicia, presbicia
 - ▣ **PRESCRIPCION** de lentes: miope, hipermétrope
 - ▣ **ASTIGMATISMO**: si, no
 - ▣ Tasa de producción de **LAGRIMAS**: reducida, normal.
 - ▣ **DIAGNOSTICO**
 - el paciente debe usar lentes de contacto duras
 - el paciente debe usar lentes de contacto blandas
 - el paciente no debe usar lentes de contacto.

Ejemplo: Prescripción de lentes de contacto

Id	Edad	Espectativa	Astigmatismo	Lagrimas	Diagnostico
1	Joven	Hipermetropía	NO	Normal	Lentes_Blandos
2	Joven	Miopía	NO	Normal	Lentes_Blandos
3	Joven	Hipermetropía	SI	Normal	Lentes_Duros
4	Joven	Miopía	SI	Normal	Lentes_Duros
5	Joven	Hipermetropía	NO	Reducida	No_usar_Lentes
...
...
22	Presbicia	Miopía	NO	Reducida	No_usar_Lentes
23	Presbicia	Miopía	NO	Normal	No_usar_Lentes
24	Presbicia	Miopía	SI	Reducida	No_usar_Lentes

<https://archive.ics.uci.edu/ml/datasets/Lenses>

Análisis de los datos disponibles

□ Tipos de Variables

- Cuantitativas y cualitativas

□ Descripciones estadísticas

- Medidas de tendencia central
- Medidas de dispersión

□ Gráficos

- Diagrama de barras
- Diagrama de torta
- Histograma
- Diagrama de caja
- Diagrama de dispersión

Tipos de variables

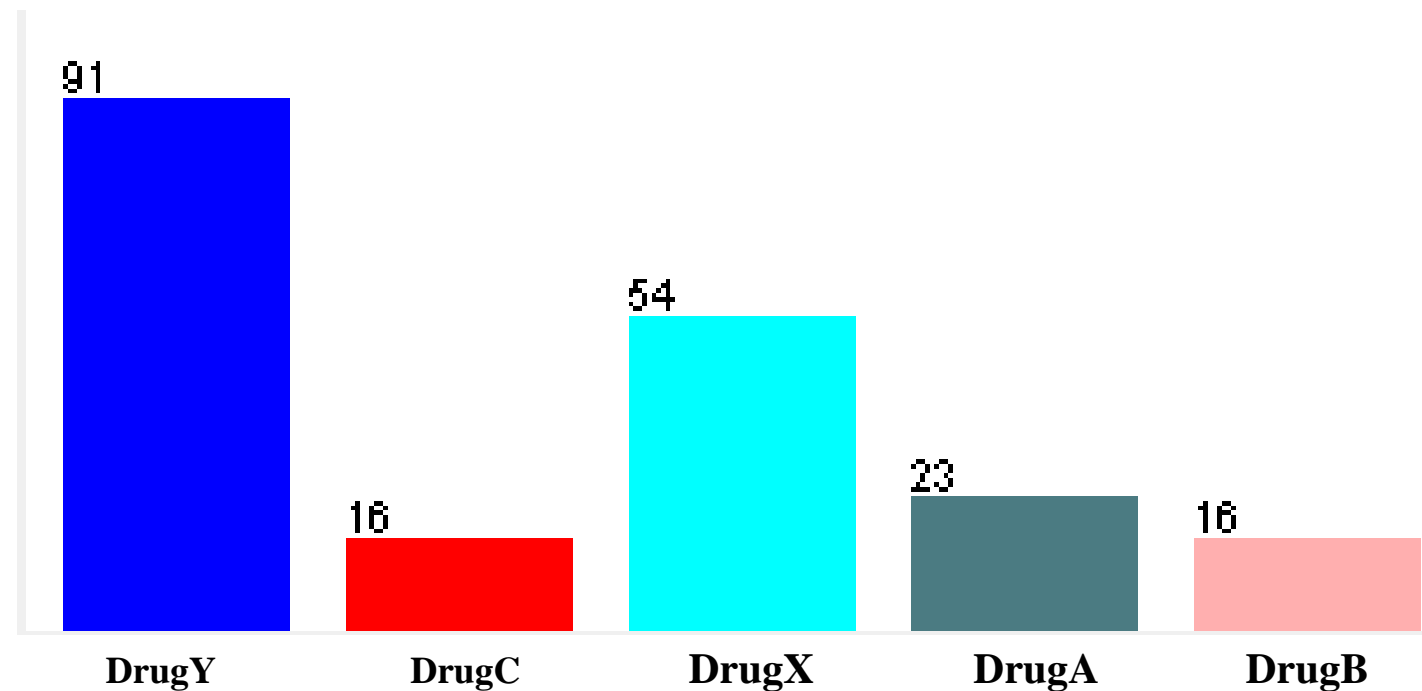
□ **Cuantitativas o numéricas**

- ▣ DISCRETAS (cant. de empleados, cant. de alumnos, etc)
- ▣ CONTINUAS (sueldo, metros cuadrados, beneficios, etc)

□ **Cualitativas o categóricas**

- ▣ NOMINALES: nombran al objeto al que se refieren sin poder establecer un orden (estado civil, raza, idioma, etc.)
- ▣ ORDINALES: se puede establecer un orden entre sus valores (alto, medio, bajo, etc)

- Se busca predecir si el tipo de fármaco que se debe administrar a un paciente afectado de rinitis alérgica es el habitual o no.



- Se dispone de información de pacientes afectados de rinitis alérgica:
 - ▣ Age: Edad
 - ▣ Sex: Sexo
 - ▣ BP (Blood Pressure): Tensión sanguínea.
 - ▣ Cholesterol: nivel de colesterol.
 - ▣ Na: Nivel de sodio en la sangre.
 - ▣ K: Nivel de potasio en la sangre.
 - ▣ Cada paciente ha sido medicado con un único fármaco de entre cinco posibles: DrugA, DrugB, DrugC, DrugX, DrugY.

Ejemplo

DRUG5.CSV

- Drug5.csv contiene 200 muestras de pacientes atendidos previamente

Nro.	Age	Sex	BP	Colesterol	Na	K	Drug
1	23	F	HIGH	HIGH	0,792535	0,031258	drugY
2	47	M	LOW	HIGH	0,739309	0,056468	drugC
3	47	M	LOW	HIGH	0,697269	0,068944	drugC
4	28	F	NORMAL	HIGH	0,563682	0,072289	drugX
5	61	F	LOW	HIGH	0,559294	0,030998	drugY
...
...
...
197	16	M	LOW	HIGH	0,743021	0,061886	drugC
198	52	M	NORMAL	HIGH	0,549945	0,055581	drugX
199	23	M	NORMAL	NORMAL	0,78452	0,055959	drugX
200	40	F	LOW	NORMAL	0,683503	0,060226	drugX

Ejemplo

Leer_Drug5.ipynb

- Drug5.csv contiene 200 muestras de pacientes atendidos previamente

Nro.	Age	Sex	BP	Colesterol	Na	K	Drug
1	23	F	HIGH	HIGH	0,792535	0,031258	drugY
2	47	M	LOW	HIGH	0,739309	0,056468	drugC
3	47	M	LOW	HIGH	0,697269	0,068944	drugC
4	28	F	NORMAL	HIGH	0,563682	0,072289	drugX
5	61	F	LOW	HIGH	0,559294	0,030998	drugY
...

- ¿Cuántos atributos tiene la tabla?
- ¿De qué tipo es cada uno de ellos?

Análisis de los datos disponibles

□ Tipos de Variables

- ▣ Cuantitativas y cualitativas

□ Descripciones estadísticas

- ▣ Medidas de tendencia central
- ▣ Medidas de dispersión

□ Gráficos

- ▣ Diagrama de barras
- ▣ Diagrama de torta
- ▣ Histograma
- ▣ Diagrama de caja
- ▣ Diagrama de dispersión

Descripciones estadísticas básicas

- Identifican propiedades de los datos y destacan qué valores deben tratarse como ruido o valores atípicos

MEDIDAS DE TENDENCIA CENTRAL

- Media
- Mediana
- Moda
- Rango medio

MEDIDAS DE DISPERSION

- Varianza
- Desviación estándar
- Rango
- Cuartiles
- Rango Intercuartil

MEDIA

- La **MEDIA** es el promedio de los valores del atributo. Dicho atributo debe ser numérico.

$$\bar{X} = \frac{\sum_{i=1}^N x_i}{N}$$

N es la cantidad de valores a promediar

- Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

$$\bar{X} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 60 + 63 + 70 + 70 + 110}{12} = \frac{696}{12} = 58$$

MEDIA

- La **MEDIA** es el promedio de los valores del atributo. Dicho atributo debe ser numérico.

$$\bar{X} = \frac{\sum_{i=1}^N x_i}{N}$$

N es la cantidad de valores a promediar

- Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

↑
 $\bar{X} = 58$

MEDIA TRUNCADA
¿cómo se calcula?
¿para qué sirve?

MEDIANA

- Divide a los valores del atributo en dos partes iguales de manera que los anteriores son todos menores que él y los siguientes son mayores.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo numérico con una **cantidad impar** de valores

30 36 47 50 52 52 56 57 60 63 70 70 110



$$\tilde{X} = x_{(N+1)/2} = 56$$

MEDIANA

- Divide a los valores del atributo en dos partes iguales de manera que los anteriores son todos menores que él y los siguientes son mayores.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo numérico con una **cantidad impar** de valores

30 36 47 50 52 52 56 57 60 63 70 70 110



$$\tilde{X} = 56$$

MEDIANA

- Divide a los valores del atributo en dos partes iguales de manera que los anteriores son todos menores que él y los siguientes son mayores.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo numérico con una **cantidad par** de valores

30 36 47 50 52 52 56 60 63 70 70 110



$$\tilde{X} = \frac{x_{N/2} + x_{(N+1)/2}}{2} = \frac{52 + 56}{2} = 54$$

MEDIANA

- Divide a los valores del atributo en dos partes iguales de manera que los anteriores son todos menores que él y los siguientes son mayores.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo numérico con una **cantidad par** de valores

30 36 47 50 52 52 56 60 63 70 70 110



$$\tilde{X} = 54$$

MEDIANA

- También puede calcularse sobre **atributos ordinales**. En tal caso, el resultado será o bien el valor que divide al conjunto en dos partes iguales o bien se dirá que “la mediana está entre los valores ...”.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo ordinal con una **cantidad impar** de valores

chico	chico	chico	chico	medio	medio	grande	grande	grande
-------	-------	-------	-------	-------	-------	--------	--------	--------



$$\tilde{X} = \text{medio}$$

MEDIANA

- También puede calcularse sobre **atributos ordinales**. En tal caso, el resultado será o bien el valor que divide al conjunto en dos partes iguales o bien se dirá que “la mediana está entre los valores ...”.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo ordinal con una **cantidad par** de valores

chico	chico	chico	medio	medio	grande	grande	grande
-------	-------	-------	-------	-------	--------	--------	--------



$$\tilde{X} = \text{medio}$$

MEDIANA

- También puede calcularse sobre **atributos ordinales**. En tal caso, el resultado será o bien el valor que divide al conjunto en dos partes iguales o bien se dirá que “la mediana está entre los valores ...”.
- Antes de calcularla deben **ordenarse los valores** del atributo.
- Ejemplo: atributo ordinal con una **cantidad par** de valores

chico	chico	chico	chico	medio	grande	grande	grande
-------	-------	-------	-------	-------	--------	--------	--------



\tilde{X} está entre “chico” y “medio”

MODA

- La moda es el valor que aparece con mayor frecuencia. Por lo tanto, puede determinarse para atributos cualitativos y cuantitativos.
- Es posible que la mayor frecuencia corresponda a varios valores diferentes, lo que da lugar a más de una MODA.
- Los conjuntos de datos con uno, dos o tres modas se denominan unimodal, bimodal y trimodal, respectivamente.
- En general, un conjunto de datos con dos o más modas es multimodal.
- Si cada valor de los datos ocurre sólo una vez, entonces no hay moda.

MODA

- La moda es el valor que aparece con mayor frecuencia. Por lo tanto, puede determinarse para atributos cualitativos y cuantitativos.

- Ejemplo: atributo numérico

30	36	47	50	52	52	56	60	63	70	70	110
----	----	----	----	----	----	----	----	----	----	----	-----

- ▣ Hay 2 modas y sus valores son 52 y 70

- Ejemplo: atributo nominal

español	inglés	chino	inglés	chino	chino
---------	--------	-------	--------	-------	-------

- ▣ La moda es “chino” por ser el valor que aparece más veces

RANGO MEDIO

- El rango medio es fácil de calcular y también puede utilizarse para evaluar la tendencia central de un conjunto de datos numéricos.
- Es la media de los valores máximo y mínimo del conjunto.

- Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

$$\text{rango medio} = \frac{\text{maximo} + \text{minimo}}{2} = \frac{110 + 30}{2} = \frac{140}{2} = 70$$

Descripciones estadísticas básicas

- Identifican propiedades de los datos y destacan qué valores deben tratarse como ruido o valores atípicos

MEDIDAS DE TENDENCIA CENTRAL

- Media
- Mediana
- Moda
- Rango medio

MEDIDAS DE DISPERSION

- Varianza
- Desviación estándar
- Rango
- Cuartiles
- Rango Intercuartil

VARIANZA Y DESVIACION ESTANDARD

- La varianza mide la dispersión de los datos con respecto a la media.
- Valores bajos indican que las observaciones de los datos tienden a estar muy cerca de la media, mientras que valores altos indican que los datos están muy dispersos.

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \bar{x}^2$$

- La desviación estándar σ es la raíz cuadrada de la varianza

VARIANZA Y DESVIACION ESTANDARD

□ Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

VARIANZA POBLACIONAL

$$\sigma^2 = \left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \bar{x}^2 = \frac{1}{12} (30^2 + 36^2 + \dots + 110^2) - 58^2 \approx 379.17$$

$$\sigma \approx \sqrt{379.17} \approx 19.47$$

VARIANZA Y DESVIACION MUESTRAL

□ Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

VARIANZA MUESTRAL

$$S^2 = \left(\frac{1}{N-1} \sum_{i=1}^N x_i^2 \right) - \bar{x}^2 = \frac{1}{11} (30^2 + 36^2 + \dots + 110^2) - 58^2 \approx 413.64$$

$$S \approx \sqrt{413.64} \approx 20.34$$

RANGO

- El rango de un conjunto de valores numéricos es la diferencia entre los valores máximo y mínimo de dicho conjunto.

- Ejemplo

30 36 47 50 52 52 56 60 63 70 70 110

$$\text{rango} = \text{maximo} - \text{minimo} = 110 - 30 = 80$$

Cuantiles, Cuartiles y Percentiles

- Los cuantiles son valores que dividen un conjunto numérico ordenado en partes iguales. Es decir que determinan intervalos que comprenden el mismo número de valores.
- Los cuantiles más usados son los siguientes:
 - ▣ CUARTILES: dividen la distribución en cuatro partes.
 - ▣ DECILES: dividen la distribución en diez partes.
 - ▣ Centiles o PERCENTILES: dividen la distribución en cien partes.
 - *El percentil es una medida de posición usada en estadística que indica, una vez ordenados los datos de menor a mayor, el valor de la variable por debajo del cual se encuentra un porcentaje dado de observaciones en un grupo.*

CUARTILES

□ Ejemplo:

30 36 47 50 52 52 56 60 63 70 70 110



$$Q_1 = 49.25$$



$$Q_2 = 54$$



$$Q_3 = 64.75$$

CUARTILES

- Los cuartiles suelen representarse como Q1, Q2 y Q3. El 2do. cuartil o Q2 coincide con la MEDIANA.
- Usaremos $(N+1)/4$ y $3(N+1)/4$ para hallar las posiciones de Q1 y Q3 respectivamente, siendo N la cantidad de valores disponibles.
 - ▣ Si no hay parte decimal, se toma directamente el elemento.
 - ▣ Si la posición corresponde a un número con parte decimal entre el elemento i y el $i+1$, se determinará un factor realizando una **interpolación lineal**.

El cuartil será:

$$Q = x_i + (x_{i+1} - x_i) * factor$$

CUARTILES

□ Ejemplo:

30 36 47 50 52 52 56 60 63 70 70 110

- La ubicación de Q_1 es $(N+1)/4$, es decir, $(12+1)/4=13/4=3.25$
- Como no es un número entero calculamos su valor entre el 3ro y el 4to elemento.

$$Q_1 = x_3 + (x_4 - x_3) * \underbrace{factor}_{\uparrow}$$

CUARTILES – cálculo del factor

i	F_i
1	0.00
2	0.09
3	0.18
4	0.27
5	0.36
6	0.45
7	0.55
8	0.64
9	0.73
10	0.82
11	0.91
12	1.00

$$N = 12$$

$$F_i = \frac{i - 1}{N - 1}$$

CUARTILES – cálculo del factor

Ubicación de Q1

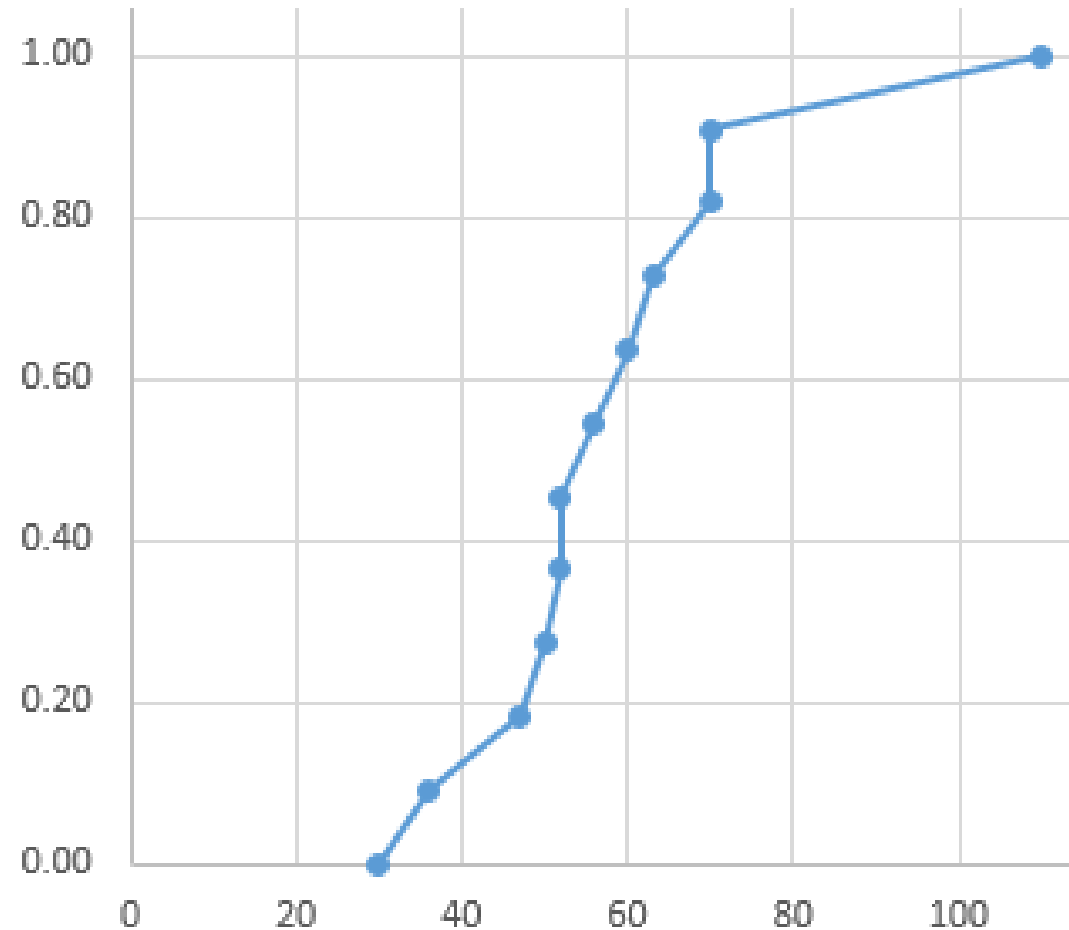
$$(N + 1)/4 = 13/4 = 3.25$$

Q1 →

X	F_i
30	0.00
36	0.09
47	0.18
50	0.27
52	0.36
52	0.45
56	0.55
60	0.64
63	0.73
70	0.82
70	0.91
110	1.00

$$N = 12$$

$$F_i = \frac{i - 1}{N - 1}$$



CUARTILES

$$F_i = \frac{i - 1}{N - 1}$$

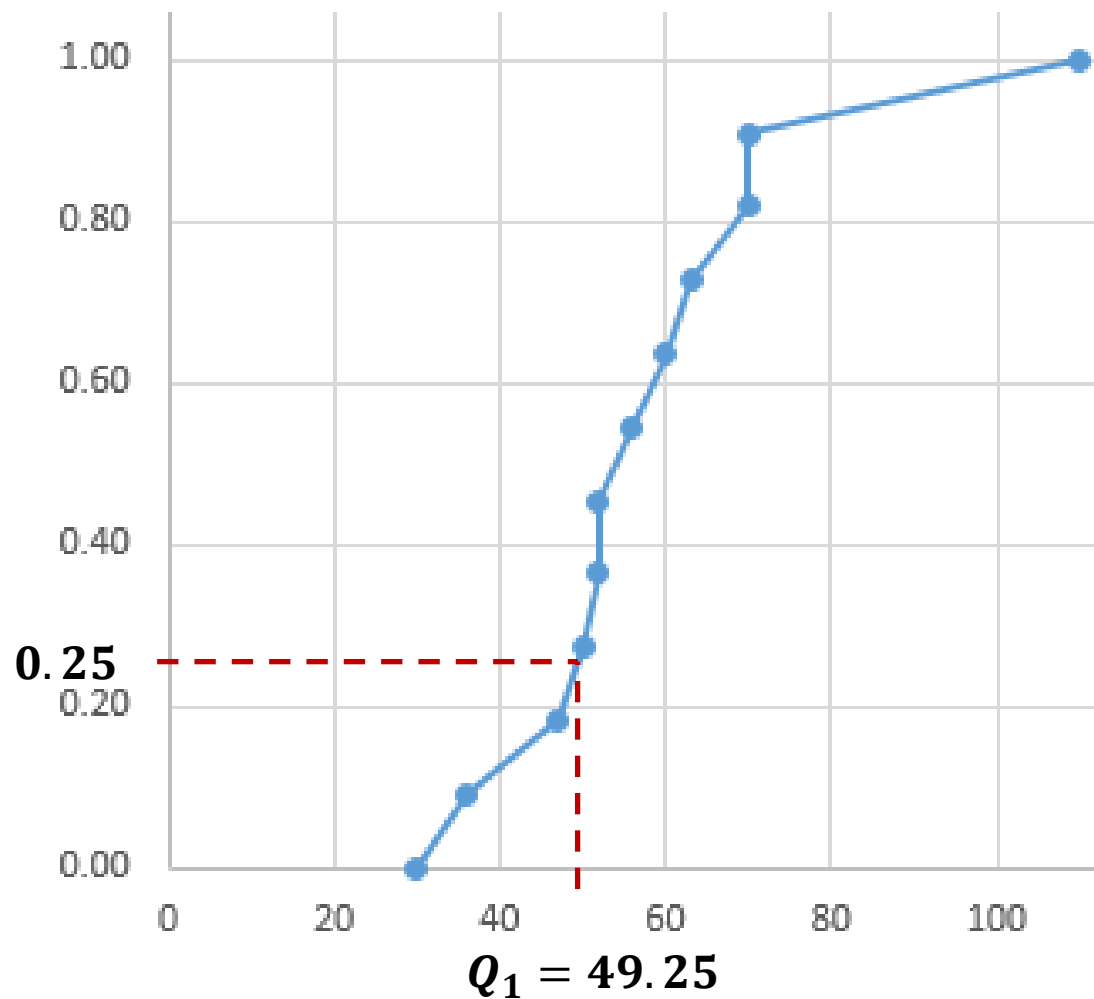
Q1



X	F_i
30	0.00
36	0.09
47	0.18
50	0.27
52	0.36
52	0.45
56	0.55
60	0.64
63	0.73
70	0.82
70	0.91
110	1.00

Ubicación de Q1

$$(N + 1)/4 = 3.25$$



CUARTILES

$$F_i = \frac{i - 1}{N - 1}$$

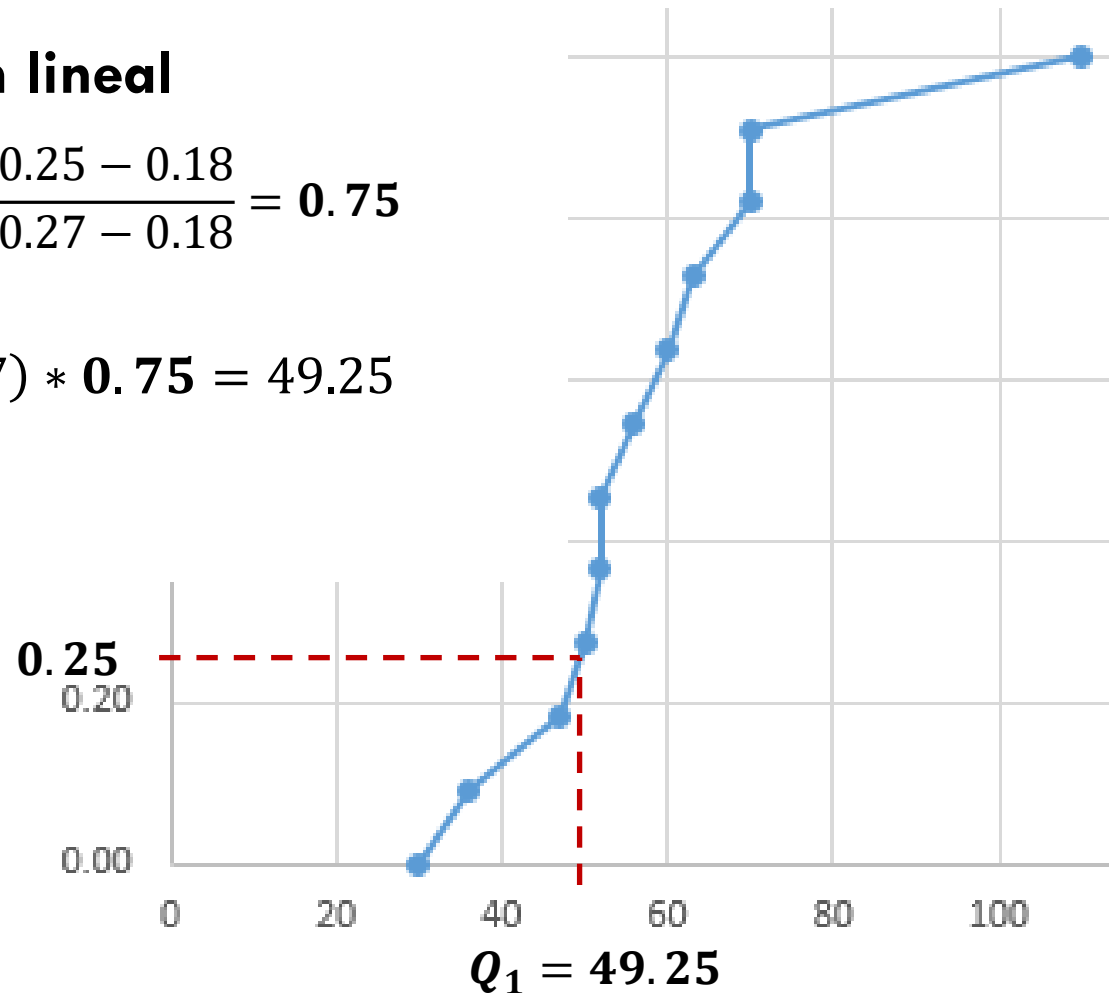
Q1 →

X	F_i
30	0.00
36	0.09
47	0.18
50	0.27
52	0.36
52	0.45
56	0.55
60	0.64
63	0.73
70	0.82
70	0.91
110	1.00

Interpolación lineal

$$factor = \frac{0.25 - F_3}{F_4 - F_3} = \frac{0.25 - 0.18}{0.27 - 0.18} = \mathbf{0.75}$$

$$Q_1 = 47 + (50 - 47) * \mathbf{0.75} = 49.25$$



CUARTILES

□ Ejemplo:

30 36 47 50 52 52 56 60 63 70 70 110

- La ubicación de Q3 es $3(N+1)/4 = 3*(12+1)/4 = 3*13/4 = 9.75$
- Como no es un número entero calculamos su valor entre el 9no y el 10mo elemento.

$$\begin{aligned} Q_3 &= x_9 + (x_{10} - x_9) * factor \\ &= 63 + (70 - 63) * 0.25 = 64.75 \end{aligned}$$

CUARTILES

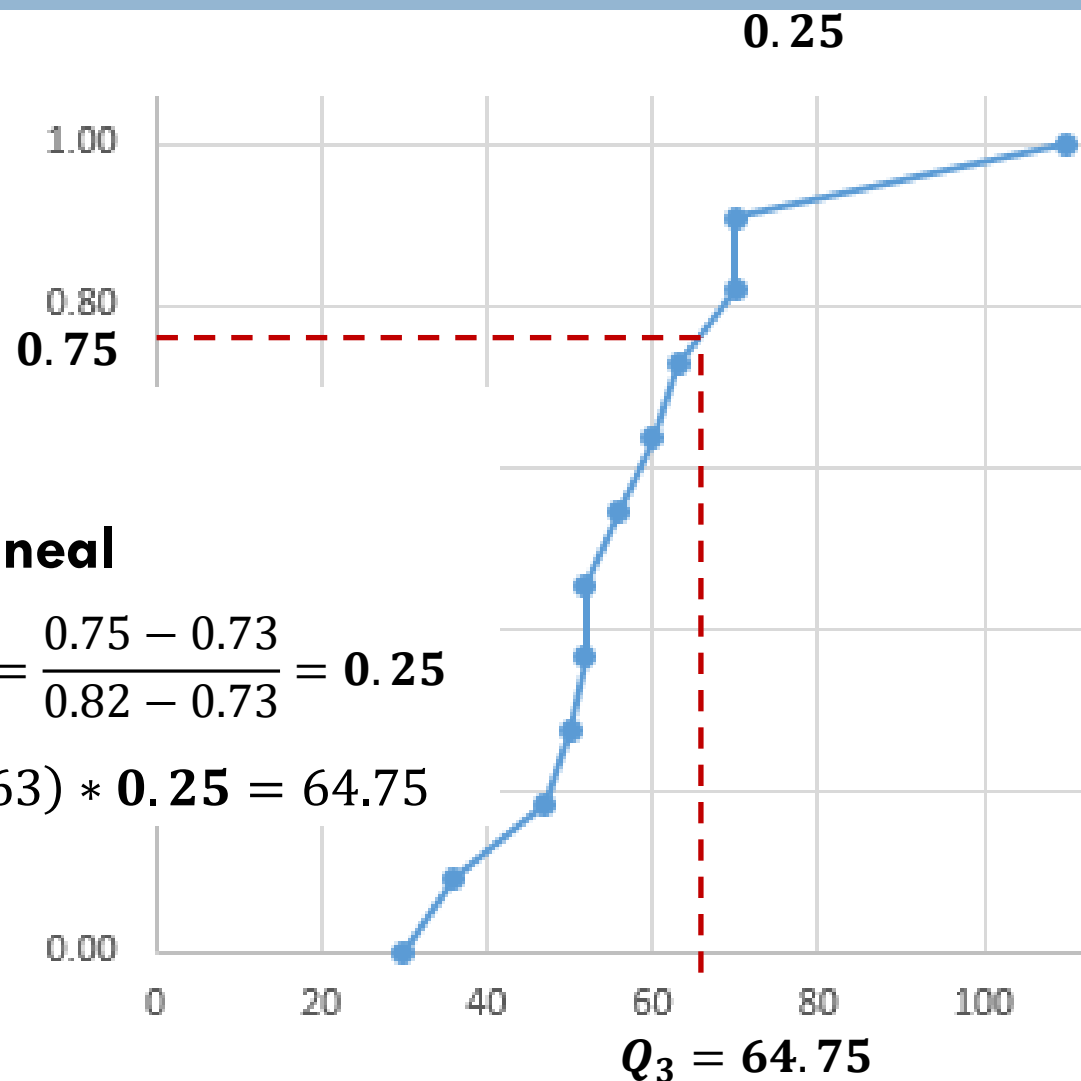
$$F_i = \frac{i - 1}{N - 1}$$

X	F_i
30	0.00
36	0.09
47	0.18
50	0.27
52	0.36
52	0.45
56	0.55
60	0.64
63	0.73
70	0.82
70	0.91
110	1.00

Interpolación lineal

$$factor = \frac{0.75 - F_9}{F_{10} - F_9} = \frac{0.75 - 0.73}{0.82 - 0.73} = 0.25$$

$$Q_3 = 63 + (70 - 63) * 0.25 = 64.75$$



CUARTILES

□ Ejemplo:

30 36 47 50 52 52 56 60 63 70 70 110



$$Q_1 = 49.25$$



$$Q_2 = 54$$



$$Q_3 = 64.75$$


RANGO INTERCUARTIL


- La distancia entre Q_1 y Q_3 es una medida sencilla de dispersión que da el rango cubierto por la mitad de los datos.
- Esta distancia se denomina **rango intercuartil (RIC)** y se define como


$$RIC = Q_3 - Q_1$$

- Ejemplo:

30 36 47 50 52 52 56 60 63 70 70 110


 $Q_1 = 49.25$


 $Q_2 = 54$


 $Q_3 = 64.75$

$$RIC = Q_3 - Q_1 = 64.75 - 49.25 = 15.50$$

Análisis de los datos disponibles

□ Tipos de Variables

- ▣ Cuantitativas y cualitativas

□ Descripciones estadísticas

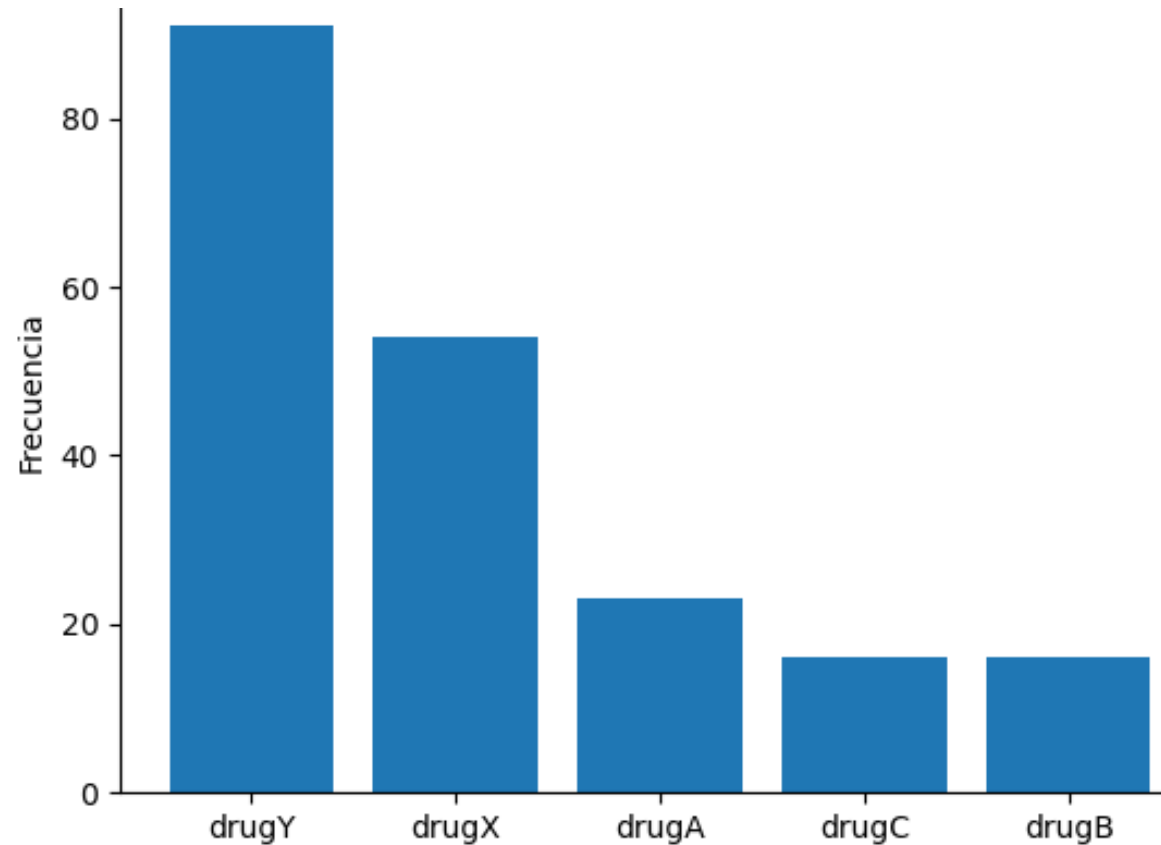
- ▣ Medidas de tendencia central
- ▣ Medidas de dispersión

□ Gráficos

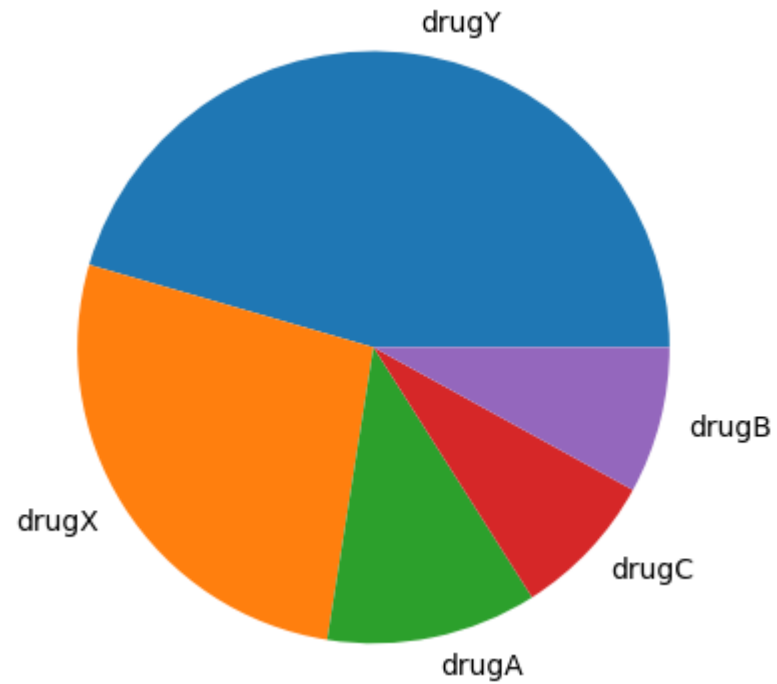
- ▣ Diagrama de barras
- ▣ Diagrama de torta
- ▣ Histograma
- ▣ Diagrama de caja
- ▣ Diagrama de dispersión

Leer_Drug5.ipynb

Atributo Drug - Diagrama de barras

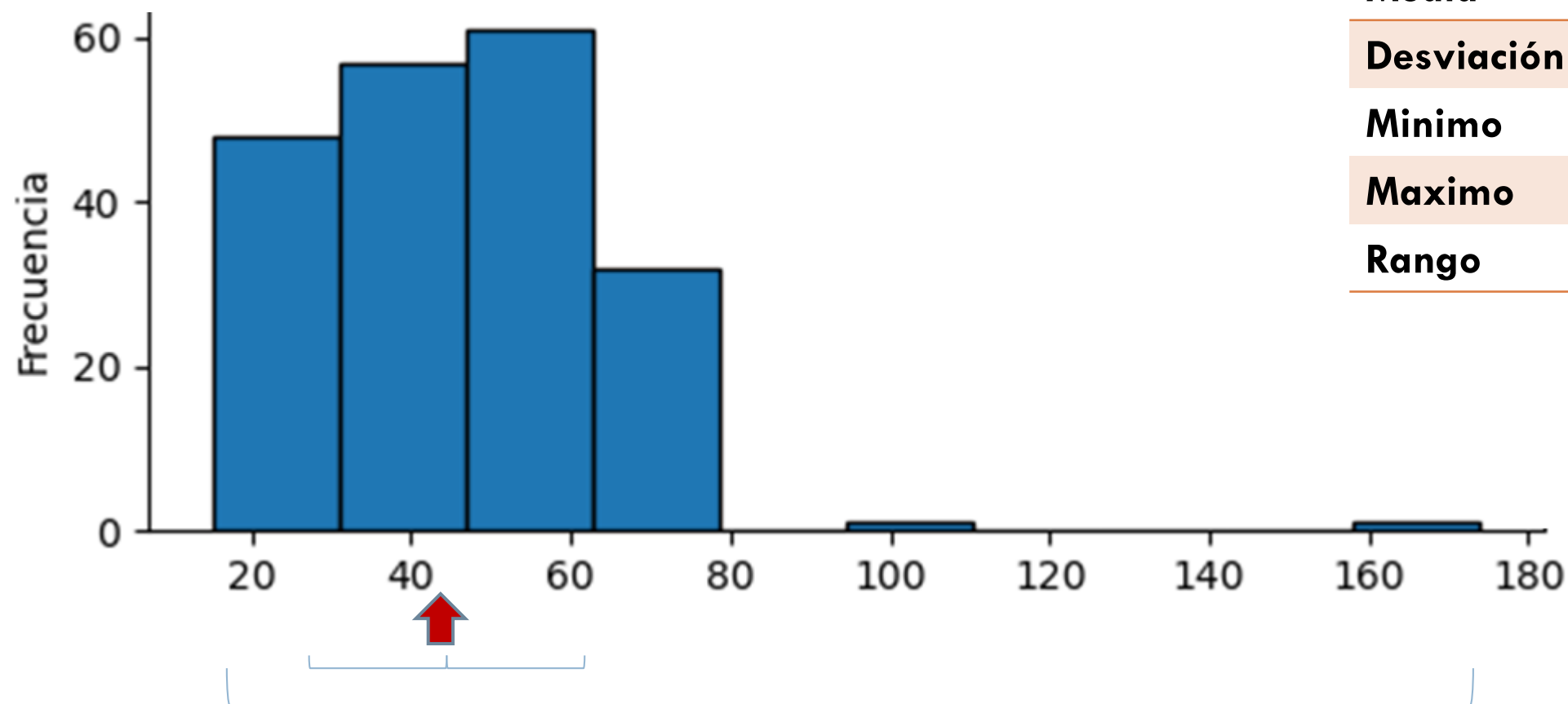


Atributo Drug - Gráfico de Torta



Atributo AGE – Histograma

(Atributo AGE del archivo Drug5_atipicos.CSV)



Media	44.965
--------------	---------------

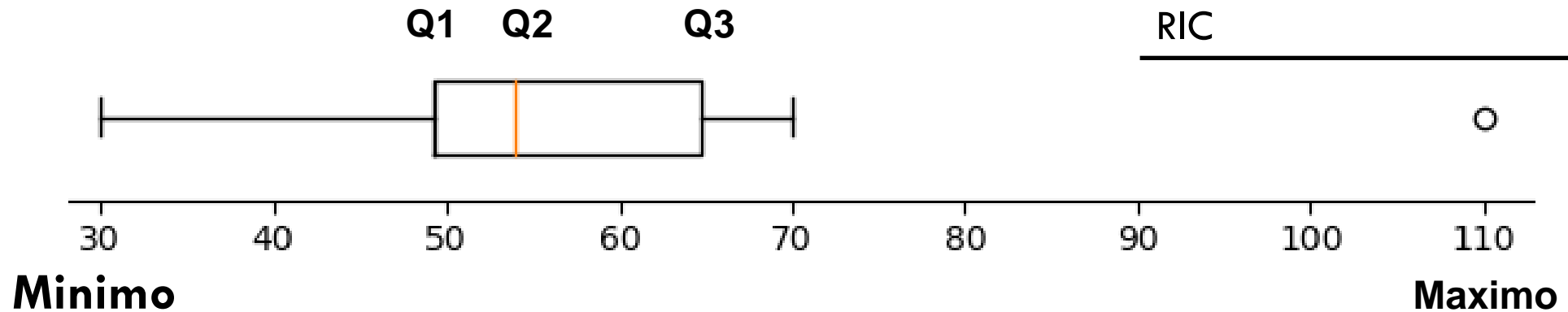
Desviación	19.145
-------------------	---------------

Minimo	15
---------------	-----------

Maximo	174
---------------	------------

Rango	159
--------------	------------

Diagrama de caja - Ejemplo



Mínimo	30
Q1	49.25
Q2 (mediana)	54
Q3	64.75
Maximo	100
RIC	

- Se consideran **valores atípicos leves** a los que se encuentran a $1.5 \times \text{RIC}$ más allá de los límites de la caja y **atípicos extremos** a los que están más allá de $3 \times \text{RIC}$.

Determine si hay valores atípicos y si son leves o extremos

Cuartiles y RIC del atributo AGE

(Atributo AGE del archivo Drug5_atipicos.CSV)

- Luego de ordenar los valores del atributo AGE deben identificarse los valores que los dividen en cuatro partes iguales.

$$Q_1 = 31$$



$$Q_2 = 45$$



$$Q_3 = 58$$



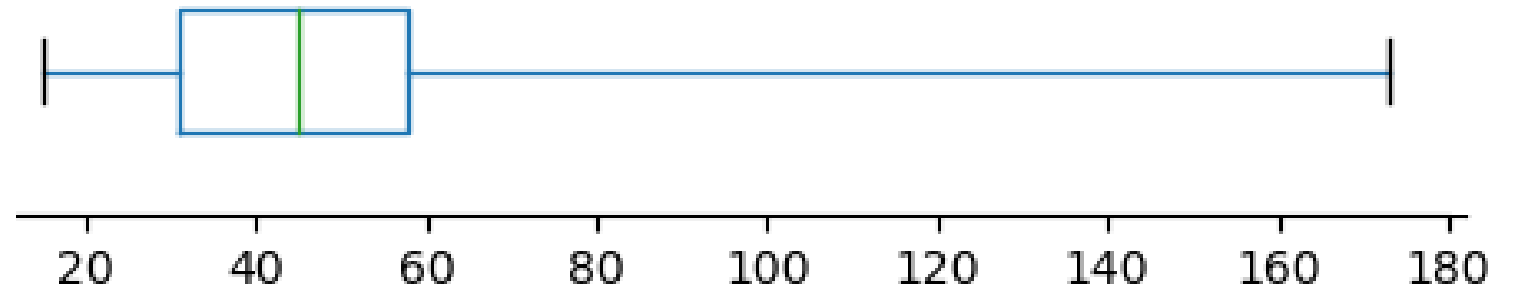
15	...	31	31	...	43	45	45	45	...	58	58	...	174
1	...	50	51	...	99	100	101	102	...	150	151	...	200

$$\text{RIC} = Q_3 - Q_1 = 58 - 31 = 27$$

Diagrama de caja (en construcción)

□ Atributo AGE (archivo Drug5_atipicos.csv)

Minimo	15
Q1	31
Q2	45
Q3	58
Maximo	174



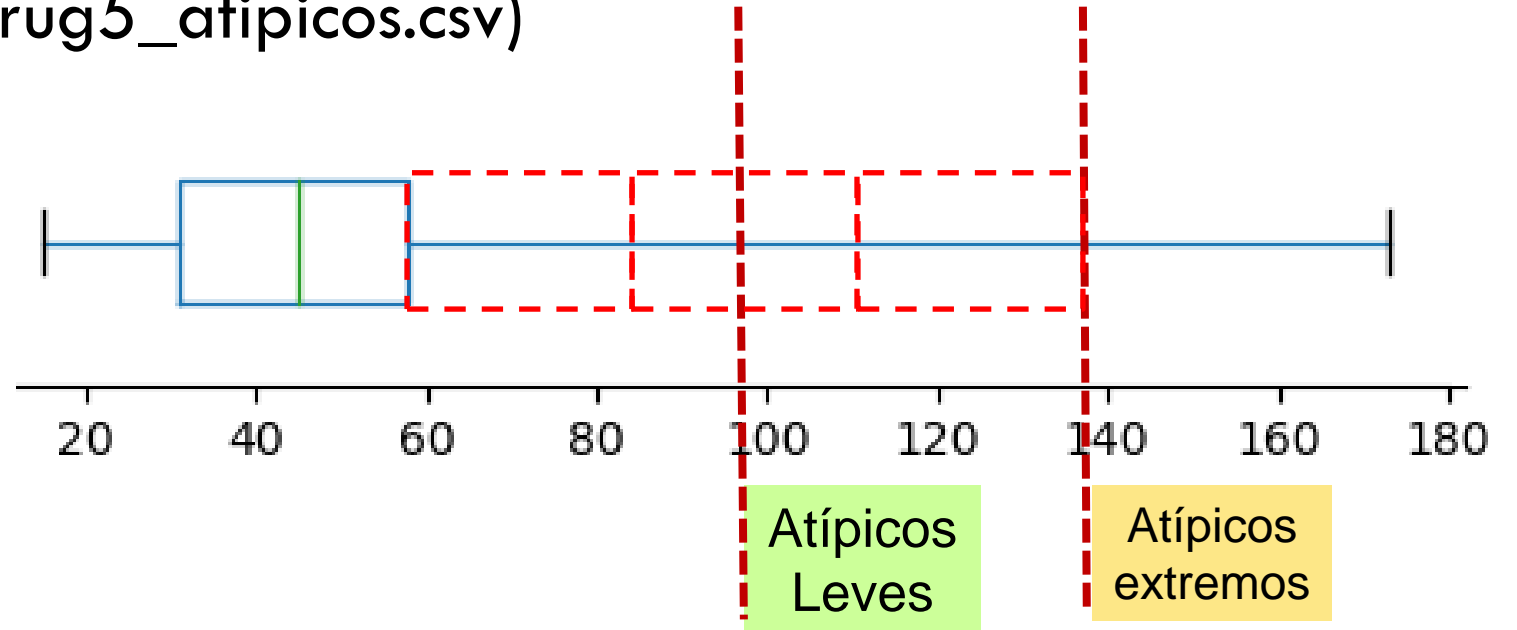
RIC	$Q3 - Q1 = 58 - 31 = 27$
Lim.Inf	$Q1 - 1.5 * RIC = 31 - 1.5 * 27 = -9.5$
Lim.Sup	$Q3 + 1.5 * RIC = 58 + 1.5 * 27 = 98.5$

Hay valores fuera de rango?

Diagrama de caja (en construcción)

□ Atributo AGE (archivo Drug5_atipicos.csv)

Minimo	15
Q1	31
Q2	45
Q3	58
Maximo	174



RIC	$Q3 - Q1 = 58 - 31 = 27$
Lim.Inf	$Q1 - 1.5 * RIC = 31 - 1.5 * 27 = -9.5$
Lim.Sup	$Q3 + 1.5 * RIC = 58 + 1.5 * 27 = 98.5$

Valor atípico o fuera de rango

- Los valores de la muestra que pertenezcan a alguno de estos intervalos

$$[Q1 - 3*RIC ; Q1 - 1.5*RIC) \text{ o } (Q3 + 1.5*RIC ; Q3 + 3*RIC]$$

serán considerados **valores fuera de rango leves**.

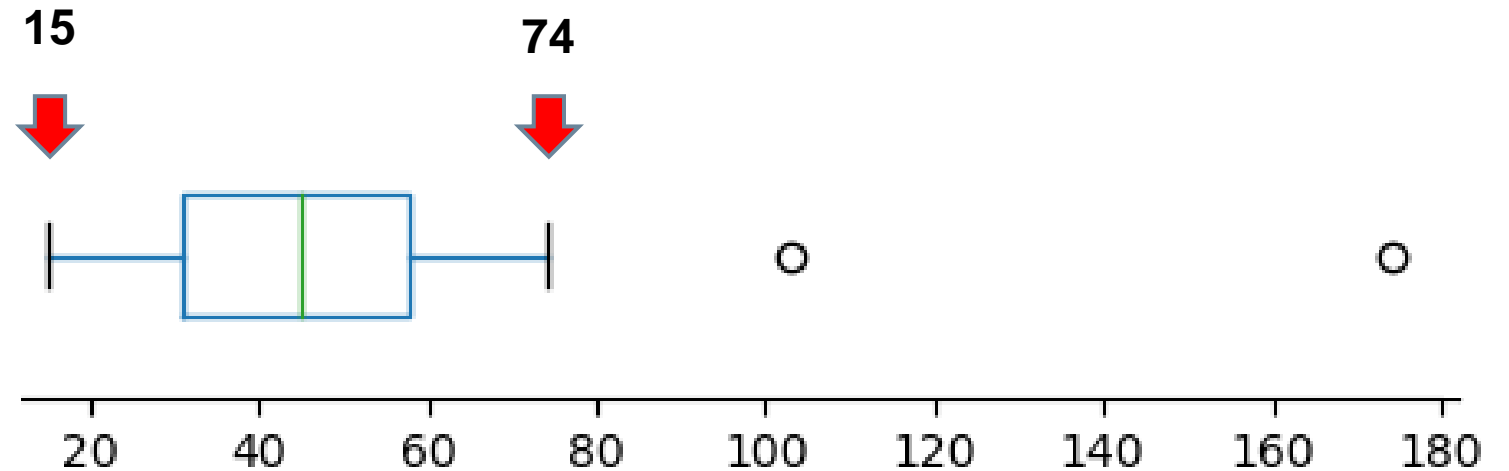
- Los valores de la muestra inferiores a

$Q1 - 3*RIC$ o superiores a **$Q3 + 3*RIC$** serán considerados **valores fuera de rango extremos**.

Diagrama de caja

□ Atributo AGE

Minimo	15
Q1	31
Q2	45
Q3	58
Maximo	174



RIC	$Q3 - Q1 = 27$
Lim.Inf	$Q1 - 1.5 * RIC = -9.5$
Lim.Sup	$Q3 + 1.5 * RIC = 98.5$

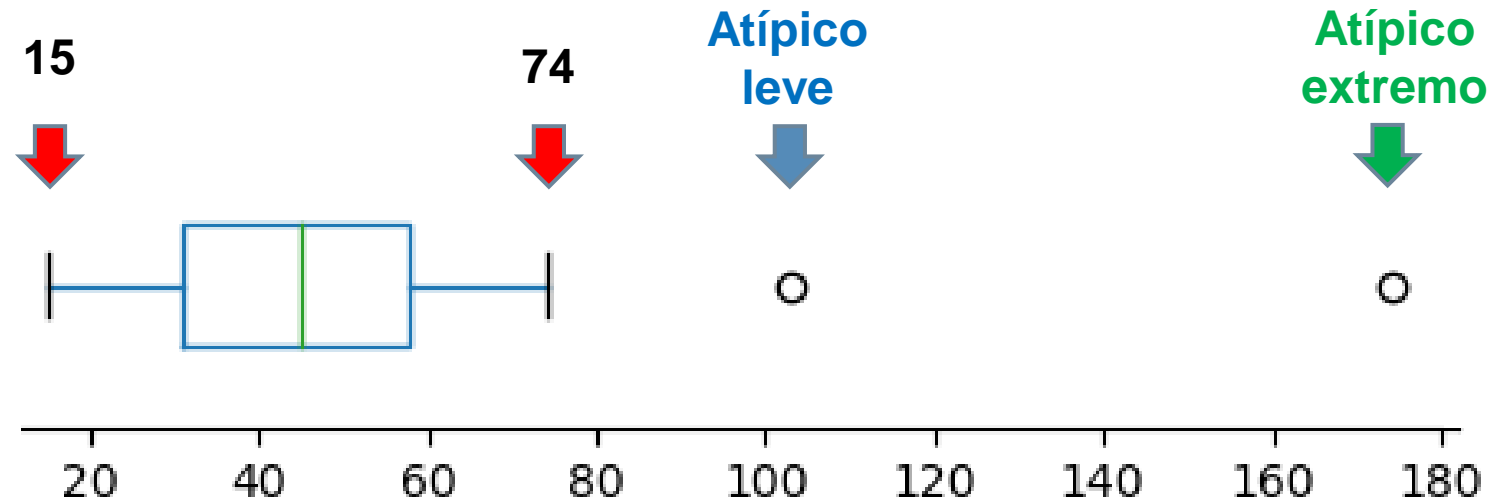
Los bigotes indican el rango de los valores de la muestra comprendidos en el intervalo

$$[Q1 - 1.5 * RIC ; Q3 + 1.5 * RIC] = [-9.5, 98.5]$$

Diagrama de caja

□ Atributo AGE

Minimo	15
Bigote Inferior	15
Q1	31
Q2	45
Q3	58
Bigote Superior	74
Maximo	174



- Los valores de AGE que pertenezcan a $[-50; -9.5)$ o $(98.5; 139]$ se considerarán **atípicos leves**.
- Los valores del atributo AGE inferiores a -50 o superiores a 139 se considerarán **atípicos extremos**.

Histograma y diagrama de caja

(Atributo AGE archivo Drug5_atipicos.CSV)

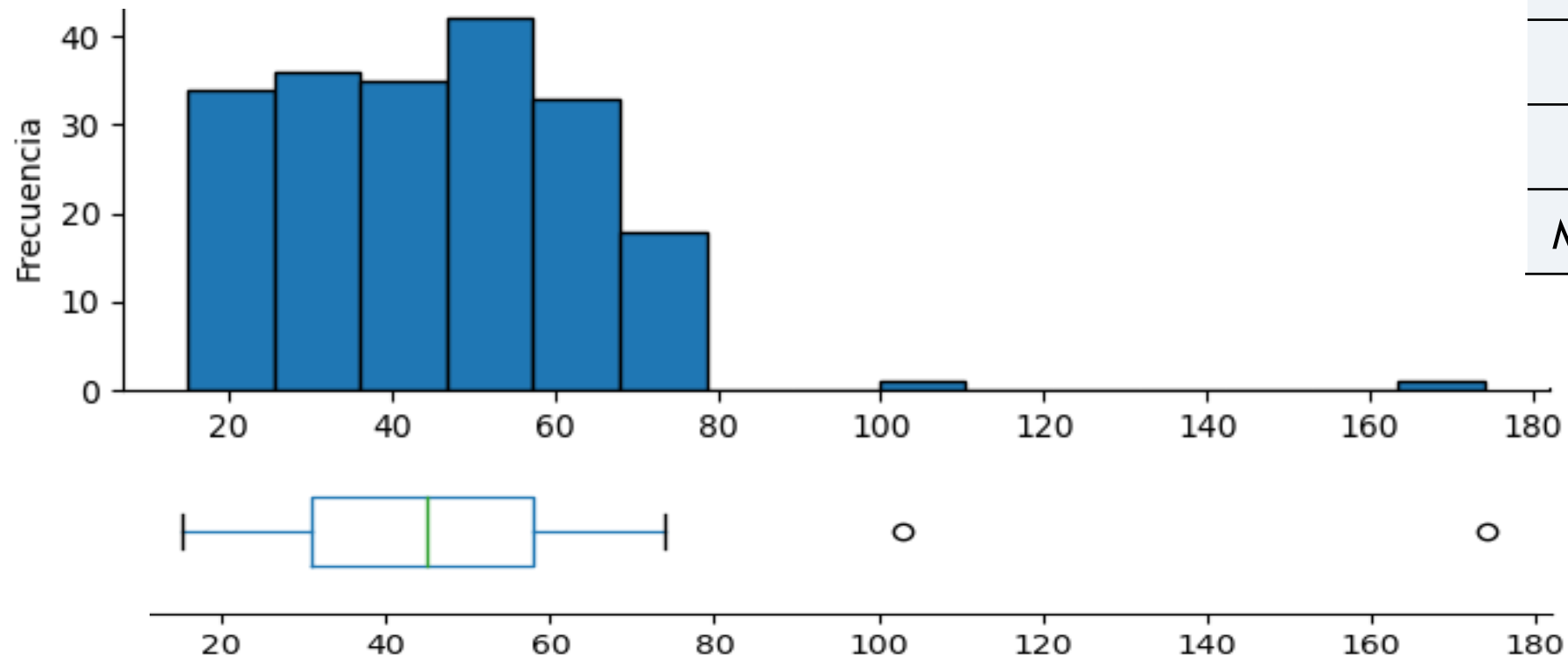
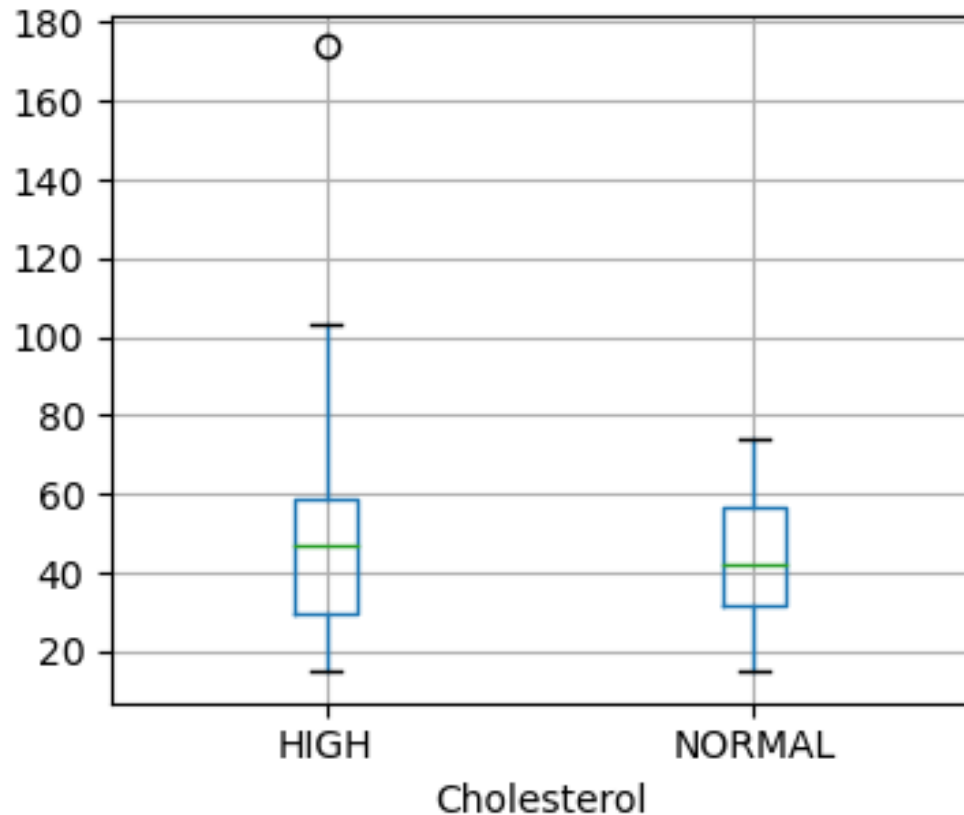


Diagrama de caja usando BY

```
df = pd.read_csv('Drug5_atipicos.csv')  
df.boxplot(column=['Age'], by='Cholesterol')
```



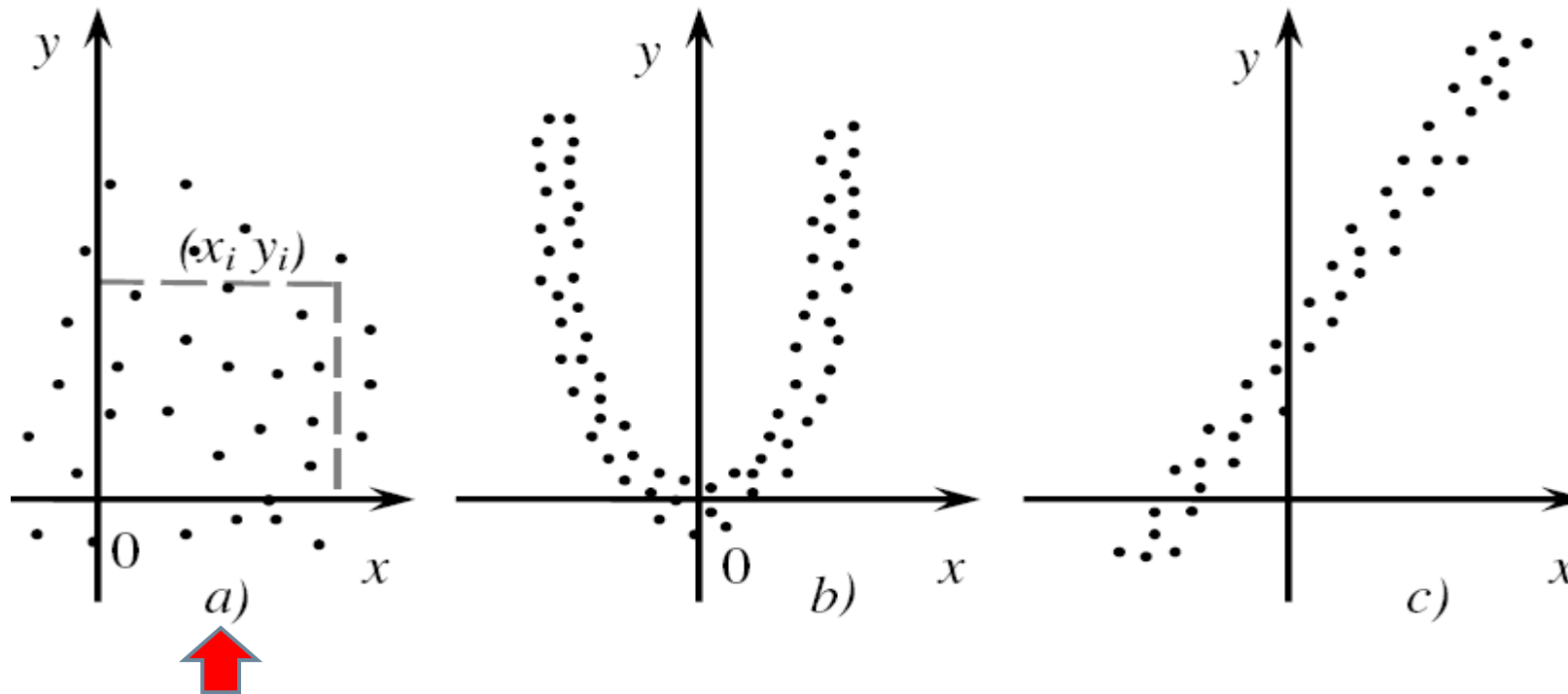
CUARTILES - Edades c/Colesterol NORMAL
[32. 42. 57.]

CUARTILES - Edades c/Colesterol HIGH
[29.5 47. 59.]

Diagrama_de_caja_agrupado.ipynb

Diagrama de Dispersión

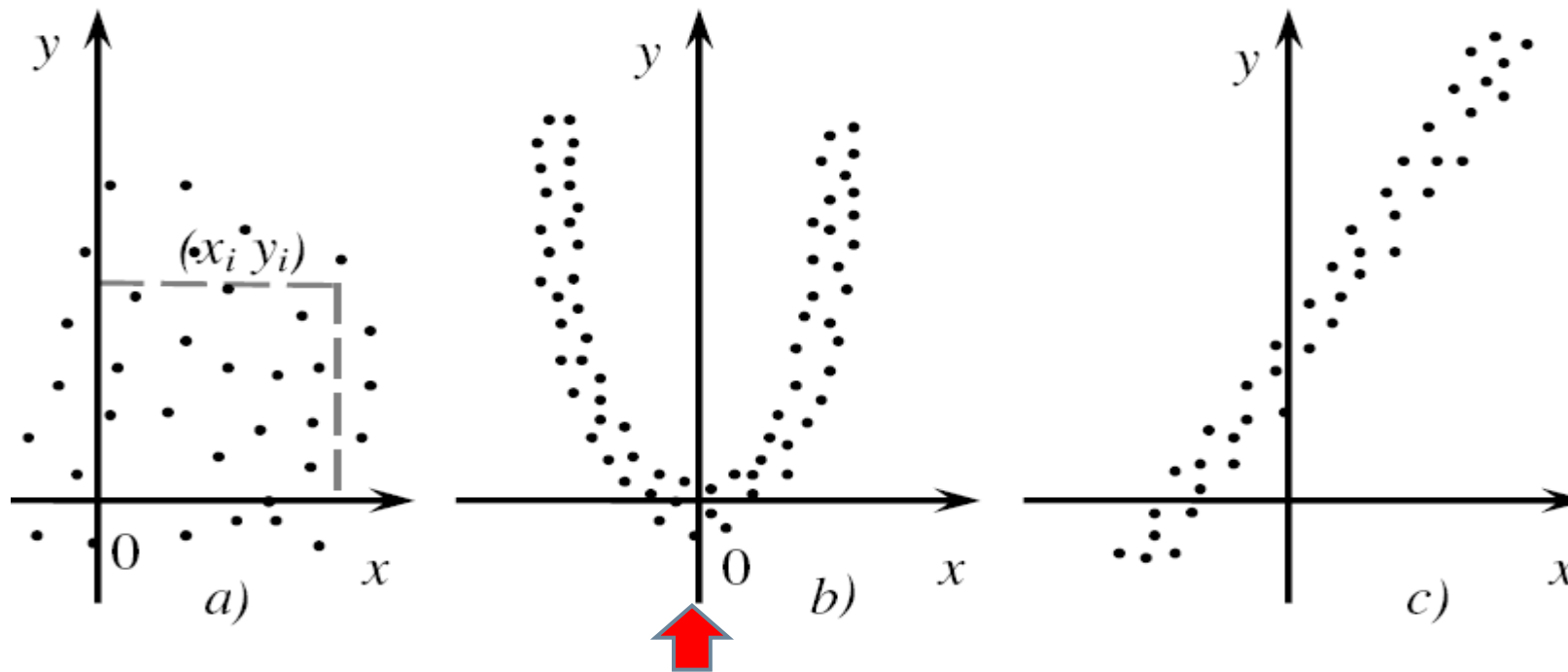
- Consiste en dibujar pares de valores (x_i, y_i) medidos de la v.a. (X,Y) en un sistema de coordenadas



Entre X e Y no hay ninguna relación funcional

Diagrama de Dispersión

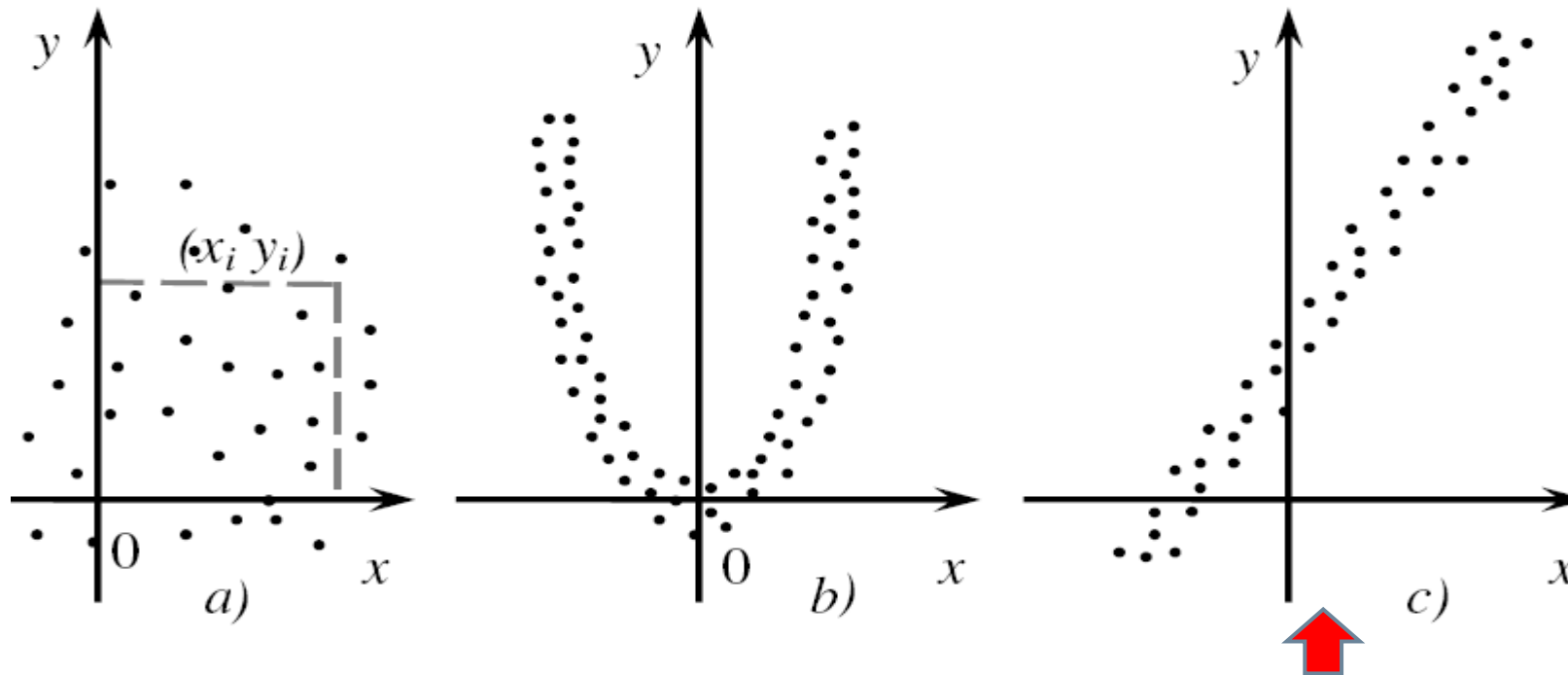
- Consiste en dibujar pares de valores (x_i, y_i) medidos de la v.a. (X,Y) en un sistema de coordenadas



Entre X e Y podría existir un relación funcional que corresponde a una parábola

Diagrama de Dispersión

- Consiste en dibujar pares de valores (x_i, y_i) medidos de la v.a. (X,Y) en un sistema de coordenadas



Entre X e Y existe una **relación lineal**. Este es el tipo de relación que nos interesa

Relación entre atributos numéricos

- Al momento de construir un modelo resulta de interés saber si dos atributos numéricos se encuentran linealmente relacionados o no. Para ello se usa el **coeficiente de correlación lineal**.

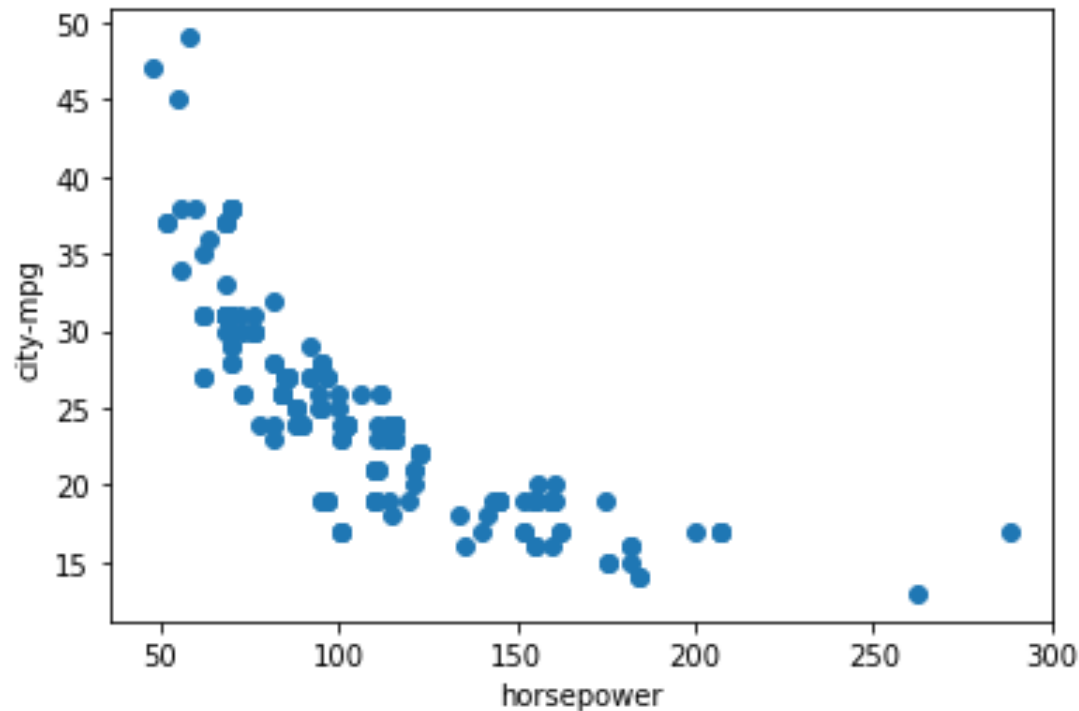


Diagrama de dispersión entre

- **Horsepower**: Potencia del motor.
- **City-mpg**: rendimiento de combustible en ciudad.

Coeficiente de correlación lineal

- Dados dos atributos X e Y el coeficiente de correlación lineal entre ellos se calcula de la siguiente forma

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

siendo $\text{Cov}(X, Y)$ la covarianza entre X e Y y σ_X y σ_Y los desvíos de cada variable.

Covarianza y desvío estándar

- Dadas dos variables X y Y

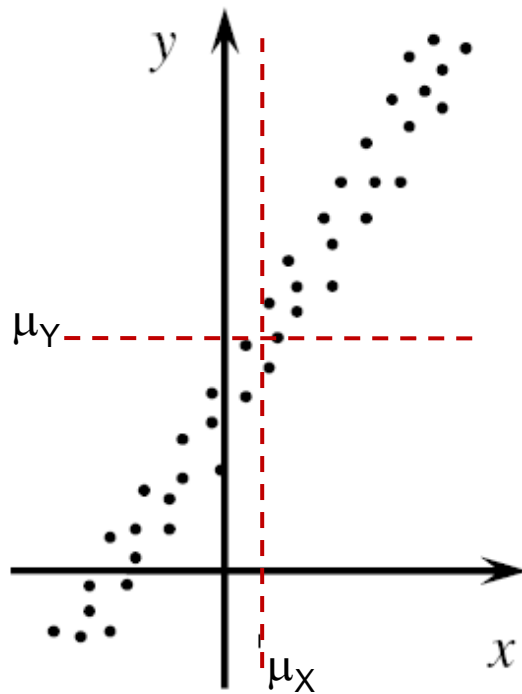
$$\text{Cov}(X, Y) = \left[\sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y) \right] / N$$

$$\sigma_X = \sqrt{\left[\sum_{I=1}^N (x_i - \mu_X)^2 \right] / N}$$

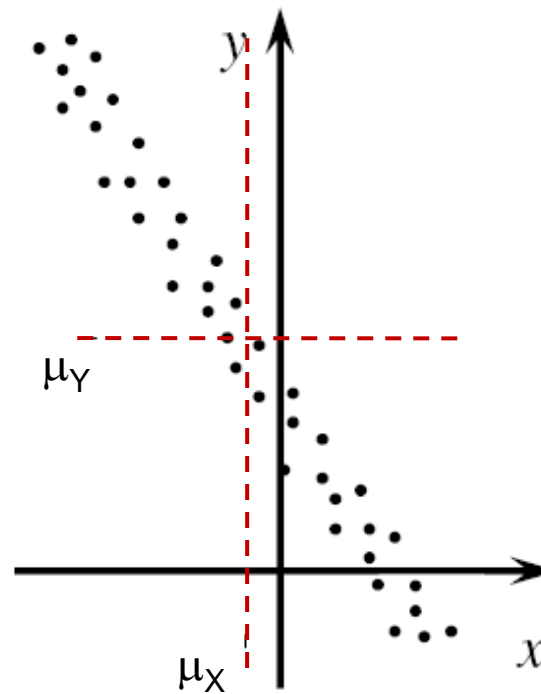
Covarianza

$$\text{Cov}(X, Y) = \left[\sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y) \right] / N$$

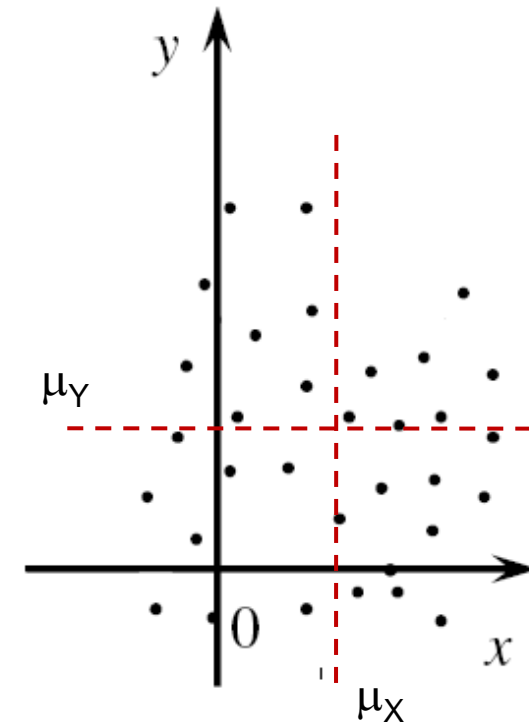
- La **covarianza** es un valor que indica el grado de variación conjunta de dos **variables aleatorias** respecto a sus medias.



Covarianza Positiva



Covarianza Negativa



Covarianza cercana a cero

Coeficiente de correlación lineal

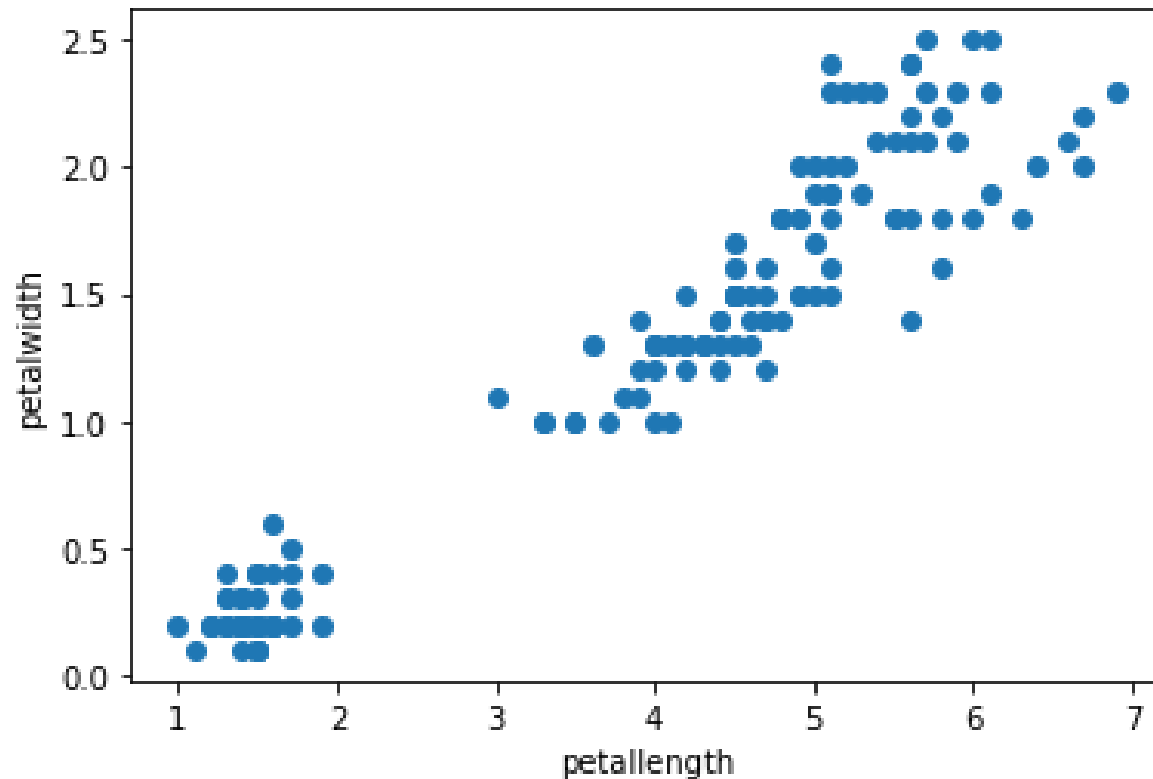
INTERPRETACION

- Si $0.5 \leq \text{abs}(\text{Corr}(A,B)) < 0.8$ se dice que A y B tienen una correlación lineal débil.
- Si $\text{abs}(\text{Corr}(A,B)) \geq 0.8$ se dice que A y B tienen una correlación lineal fuerte
- Si $\text{abs}(\text{Corr}(A,B)) < 0.5$ se dice que A y B no están correlacionados linealmente. Esto NO implica que son independientes, sólo que entre ambos no hay una correlación lineal.

Ejemplo

Correlacion_Iris.ipynb

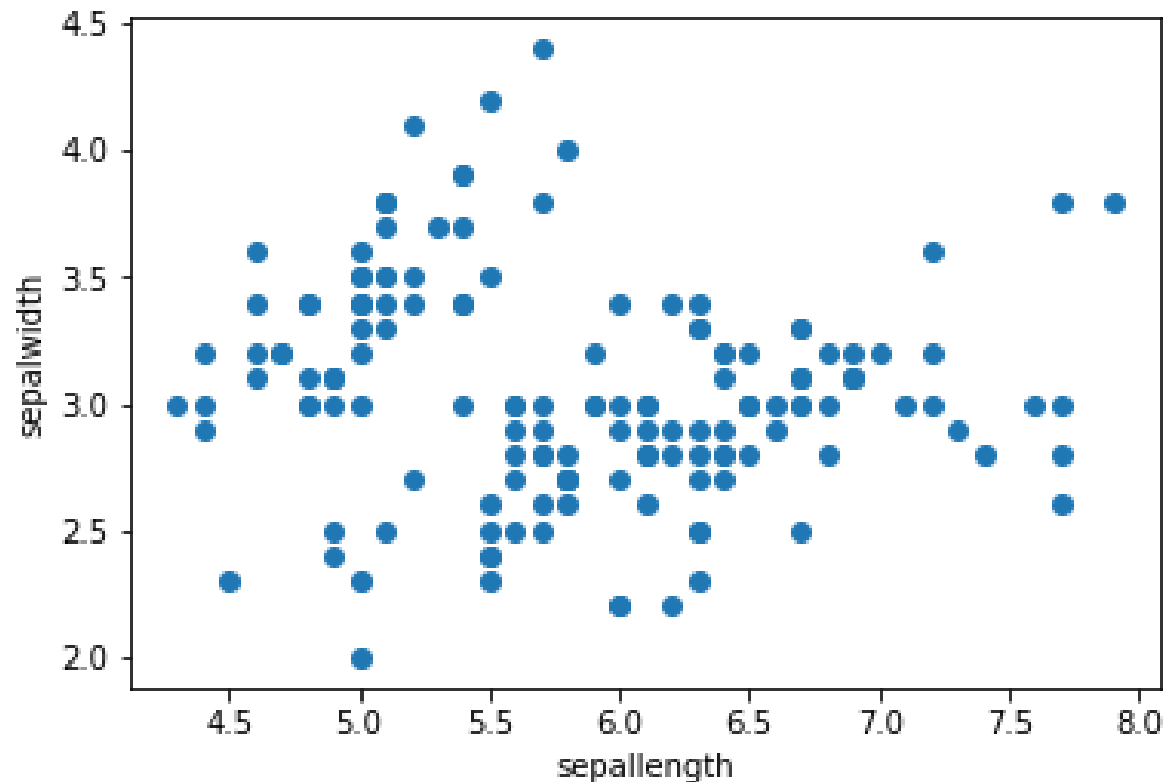
- El valor del **coeficiente de correlación lineal** entre los atributos PETALLENGTH y PETALWIDTH es **0.96**



Ejemplo

Correlacion_Iris.ipynb

- El valor del **coeficiente de correlación lineal** entre los atributos SEPALLENGTH y SEPALWIDTH es **-0.11**



Resumen

□ Tipos de aprendizaje

- ▣ Supervisado
- ▣ No supervisado

□ Redes neuronales

- ▣ Predicción
- ▣ Segmentación o agrupamiento

□ Tipos de Variables

- ▣ Cuantitativas y cualitativas

□ Descripciones estadísticas

- ▣ Medidas de tendencia central
- ▣ Medidas de dispersión

□ Gráficos

- ▣ Diagrama de barras
- ▣ Diagrama de torta
- ▣ Histograma
- ▣ Diagrama de caja
- ▣ Diagrama de dispersión
- Coeficiente de correlación lineal