Universidad del Rosario

Analisis Avanzado de Datos

W1. Introducción

FERNEY ALBERTO BELTRAN MOLINA
Escuela de Ingeniería, Ciencia y Tecnología
Matemáticas Aplicadas y Ciencias de la Computación

Profesor

FERNEY ALBERTO BELTRAN MOLINA
ferney.beltran@urosario.edu.co

Ingeniero Electrónico.
Magister en TIC
Candidato Doctor en TIC

Director del Centro de investigación e innovación CEINTECCI.
Miembro de la junta directiva Avanciencia

Procesamiento y análisis de datos basadas en IA.
Simulación y modelado por computación,
Optimizan Sistemas de procesamiento en hardware y software
Diseño de sistemas electrónicos reconfigurables

Propósito del curso

- Presenta modelos estadísticos avanzados para el análisis de datos.
- Generalizaciones al modelo de regresión lineal que permiten mayor flexibilidad de estimación.
- Introduce modelos de datos estructurados y variables latentes, así como modelos para el análisis de datos dependientes.
- Fundamentos teóricos y sus aplicaciones

Resultados de Aprendizaje

- Evaluar la bondad de un modelo estadístico.
- Emplear modelos lineales generalizados para la estimación de relaciones entre variables.
- Identificar situaciones donde el uso de métodos de suavización es adecuado.
- Estimar modelos estadísticos usando variables latentes y datos estructurados.
- Entender y utilizar los datos dependientes.

Actividades de Evaluación

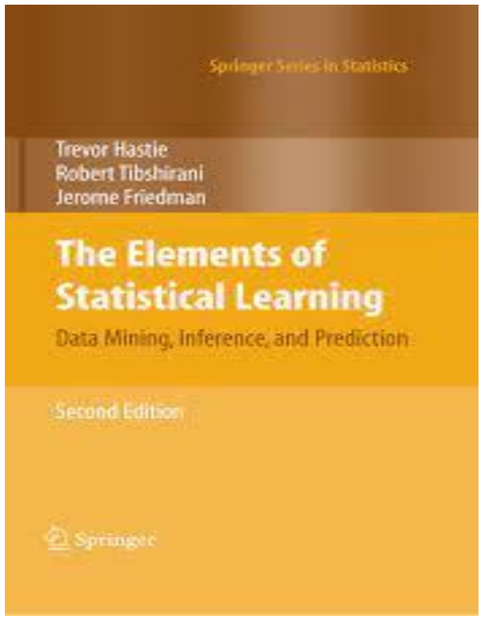
Heteroevaluación	70 %	Autoevaluación	15 %	Coevaluación	15 %

5 momentos

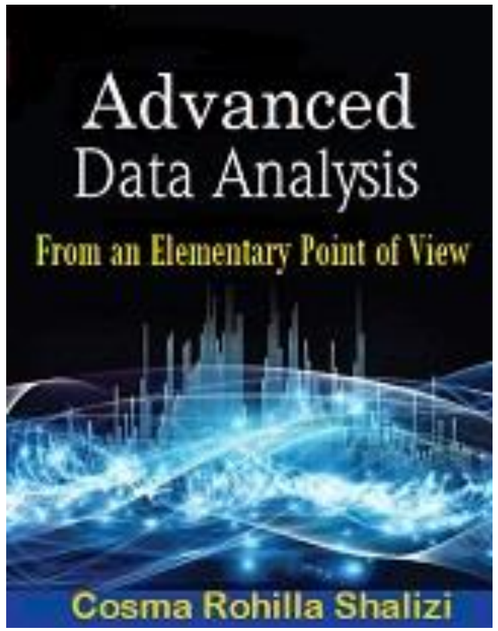
3 evaluaciones/ miniproyectos	60 %	1 evaluación general	15 %	Proyecto	25 %

31 agosto	5 de octubre	2 de Noviembre	23 de Noviembre	28 Noviembrea 1 junio
20%	20%	20%	15%	25%

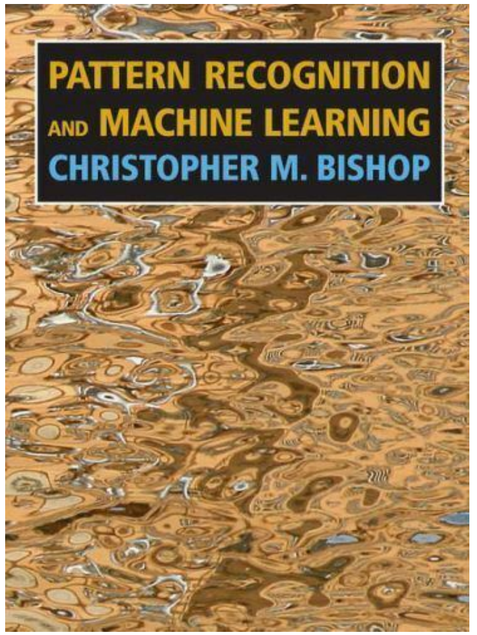
Libros guía



Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning. Springer.



Shalizi, Cosma. Advanced data analysis from an elementary point of view. 2021



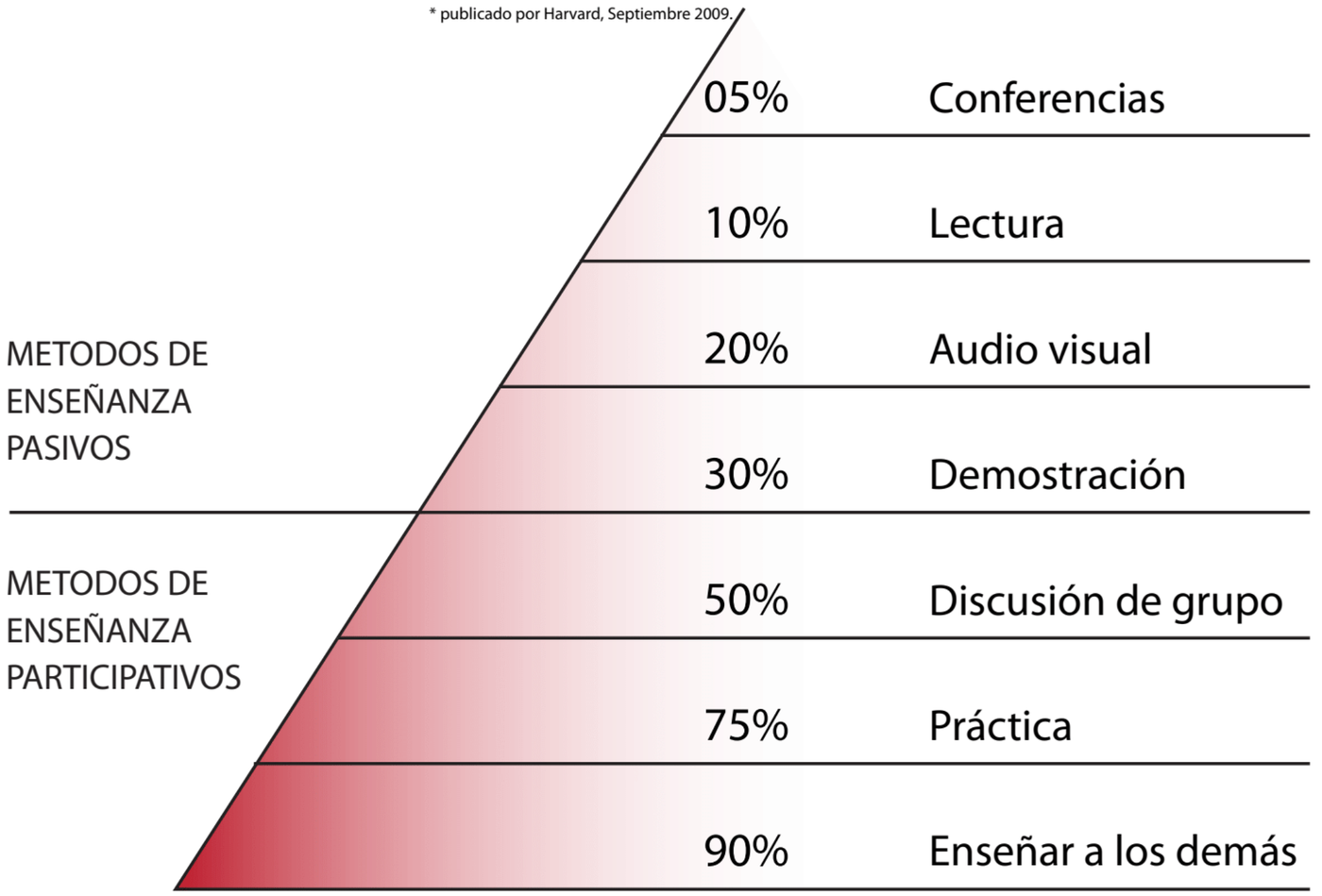
3. Bishop, Christopher M. Pattern Recognition and Machine Learning. Springer. 2006.

Y el software a usar

Usaremos **colab** y **Python**:

- Editor de texto de su preferencia

La pirámide de aprendizaje*
% de retención

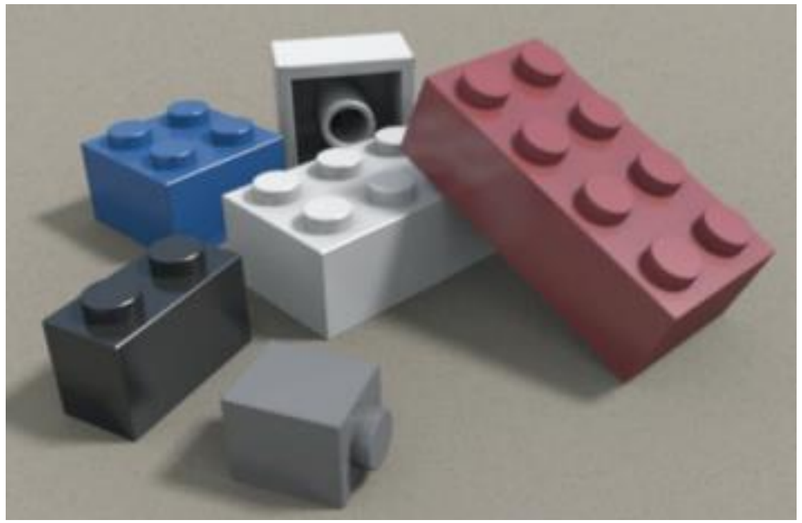


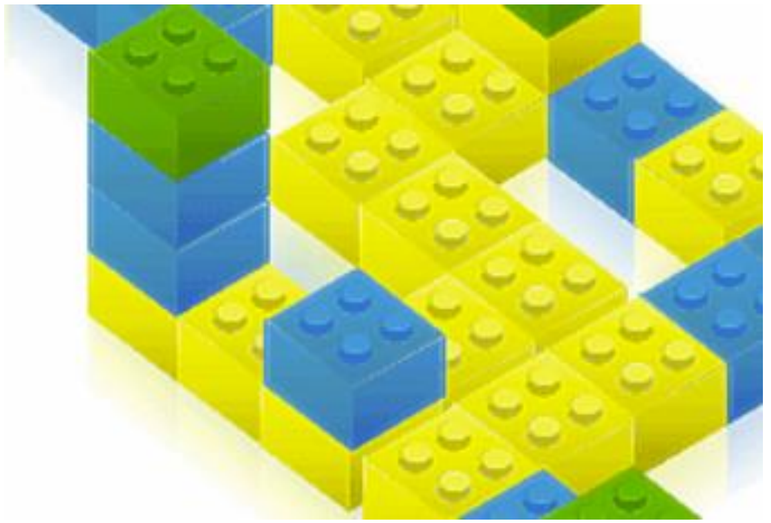
Reflexiones

¿Cuando fue la ultimas vez que hiciste algo por primera vez ?

Lego & Creativa colaboración

Cada uno recibirá una tarea. Léela varias veces para asegurarte de entenderla. No la muestres a nadie más.





Tres factores claves en tareas complejas

1. Trabajar en equipo y confiar en la dirección que se mueve el equipo
2. Nadie tiene la respuesta
3. El resultado no se da adelantado

Tres niveles de rendimiento

1. Confusión
2. Control
3. Co-creación

Reflexiones

1. Qué funcionó?
2. Qué fue un desafío ?

Alguien del equipo facilite el dialogo y usemos el tablero para realizar las anotaciones

Reflexiones

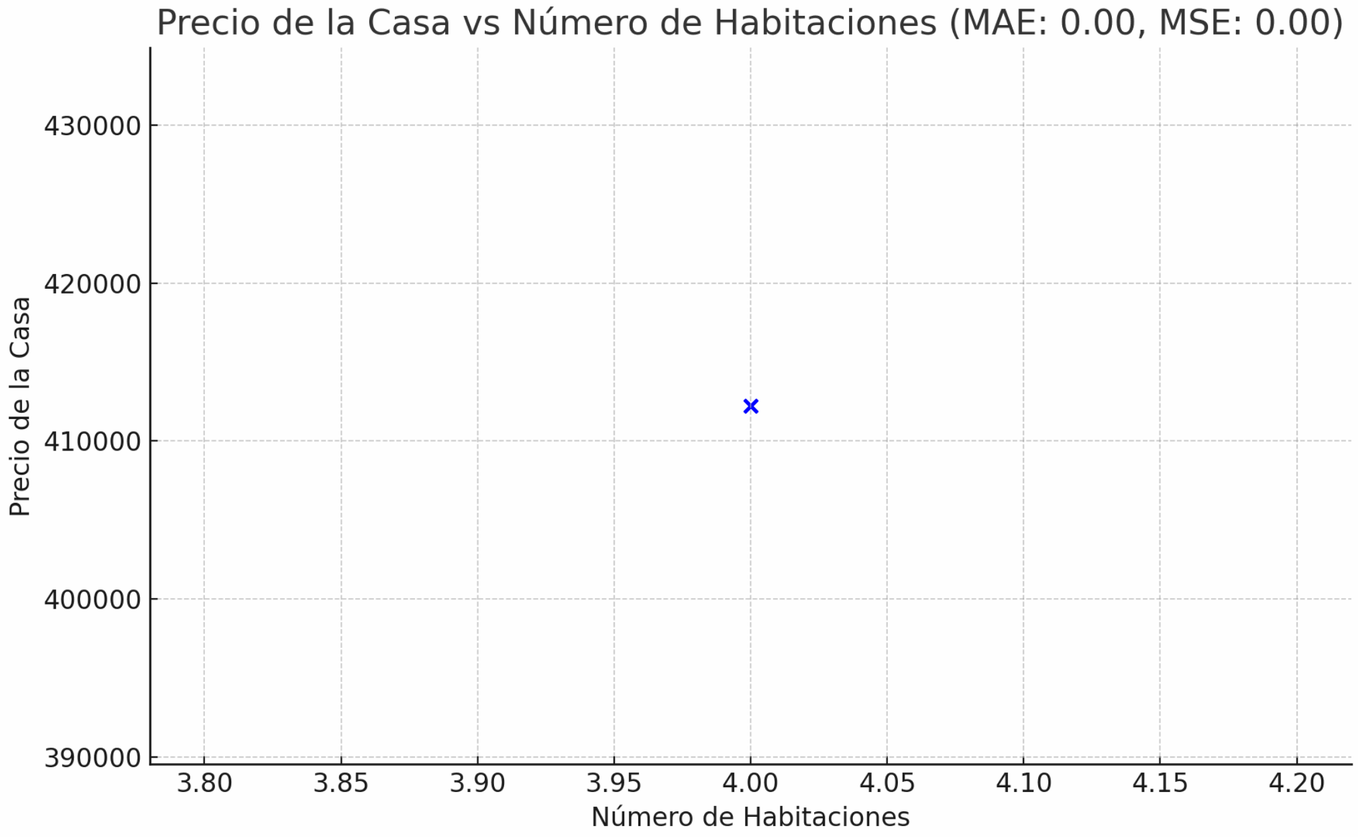
1. Cada jugador actúa como un "modelo" con restricciones y esas restricciones afectan la construcción de la torre
2. Cada jugador evalué su desempeño durante el juego
3. Liderazgo efectivo y colaboración
4. Limitaciones y flexibilidad que los analistas pueden enfrentar
5. Habilidades que adquirieron durante el juego y cómo pueden aplicar esas habilidades en el contexto del curso

Reflexiones del curso

Cuales son nuestras reglas para análisis de datos

Qué esperan de este curso?

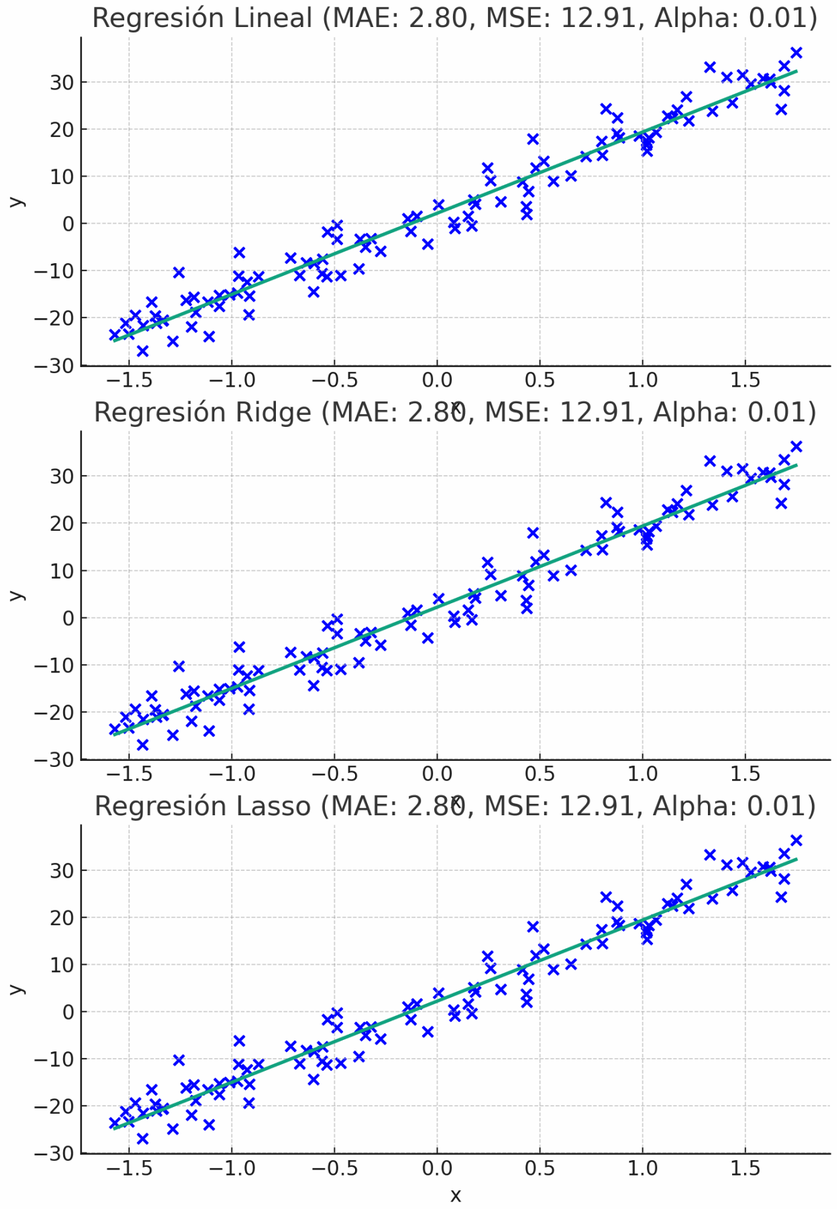
Introducción y Repaso de Regresión



- Repaso de conceptos básicos de regresión lineal

Conjunto de datos que contiene información sobre el precio de casas y diferentes características como el número de habitaciones, el área del terreno, etc.

Utilizarán regresión lineal para predecir el precio de las casas en función de estas características y evaluarán el rendimiento del modelo utilizando métricas como el error medio absoluto (MAE) y el error cuadrático medio (MSE).

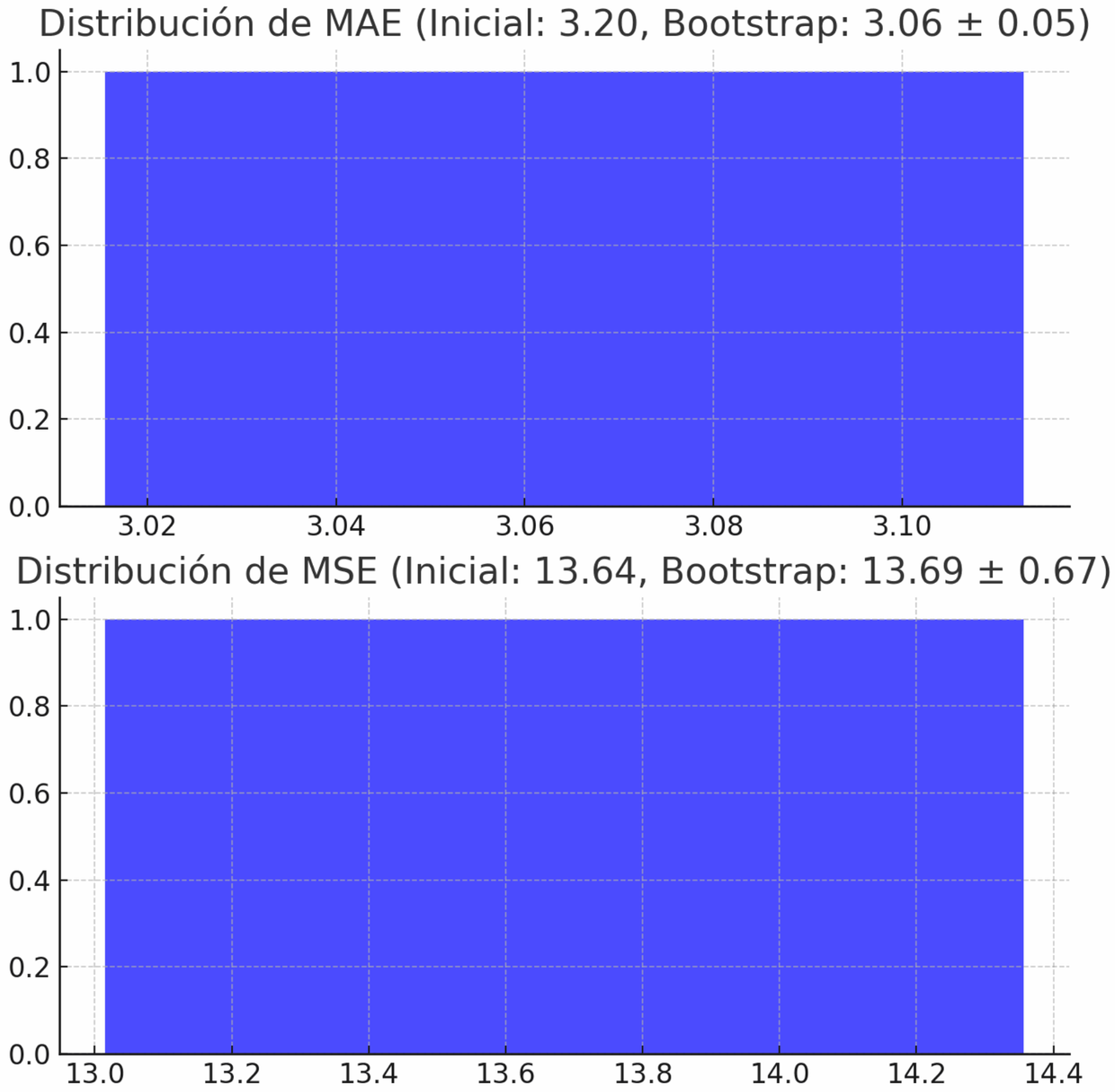


Regresión Lineal Múltiple y Métodos de Regularización

- Ridge y Lasso.
- Comprendan cómo aplicar estos métodos para mejorar la precisión de los modelos de regresión y evitar el sobreajuste.

Conjuntos de datos relacionados con el rendimiento académico de los estudiantes, incluyendo factores como horas de estudio, asistencia a clases, calificaciones previas, etc.

Se aplica regresión lineal múltiple y utilizarán los métodos de regularización (Ridge y Lasso) para mejorar la precisión del modelo y evitar problemas de sobreajuste.



Evaluación de Modelos y Métodos de Bootstrap

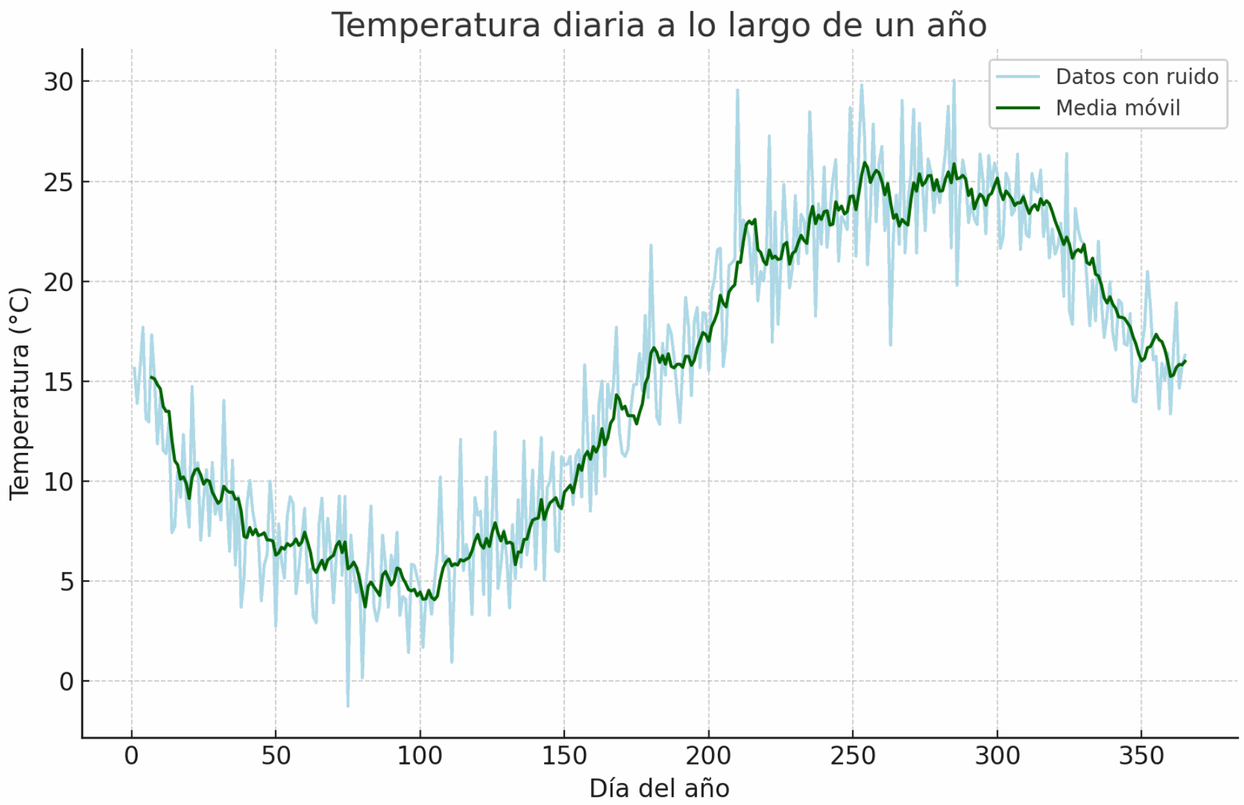
- Métricas de rendimiento para evaluación de modelos.
- Concepto y aplicación del método Bootstrap para estimar la incertidumbre en los resultados del modelo.

Aprenderán cómo evaluar la calidad de sus modelos y cómo utilizar el método Bootstrap para obtener estimaciones robustas.

Métodos de Suavización

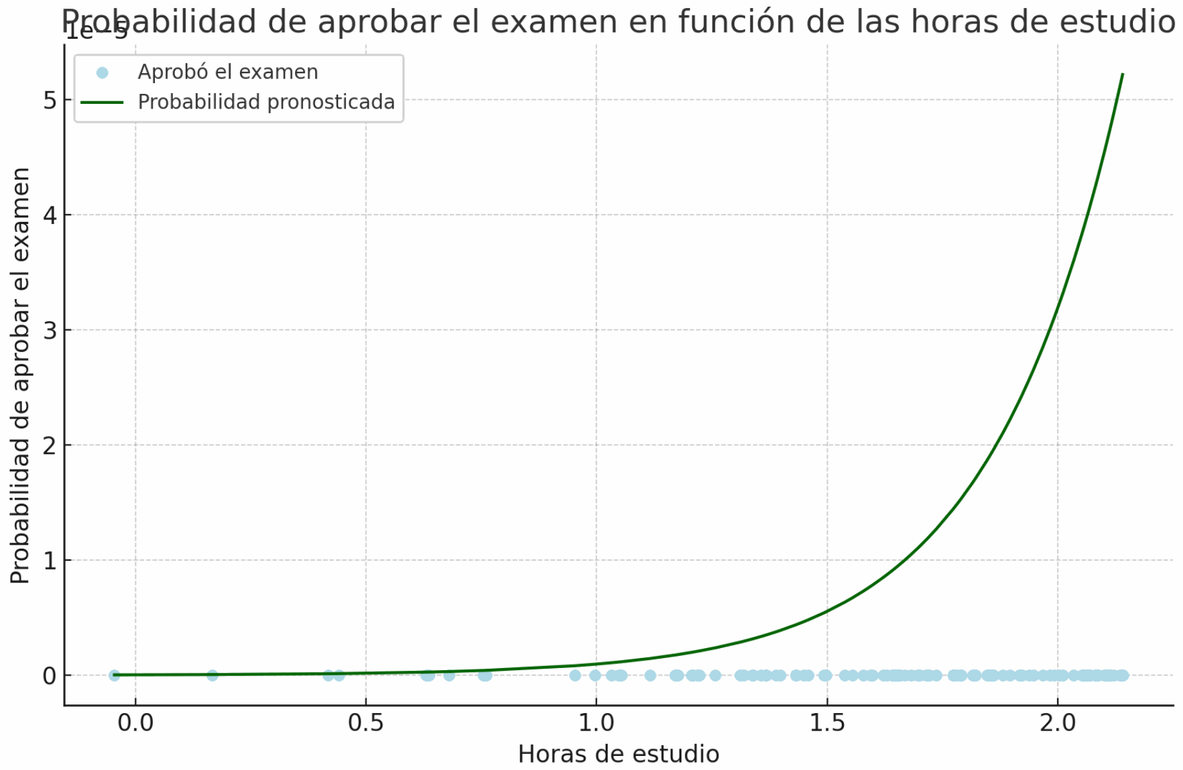
- Utilizando polinomios a trazos y Splines.
- Métodos de kernel para suavización de curvas y la estimación de densidades mediante métodos de kernel.

Aprenderemos cómo aplicar estos métodos para capturar tendencias subyacentes en datos ruidosos, a suavizar datos y obtener estimaciones de densidades más precisas.



Conjunto de datos de temperaturas diarias registradas en una ciudad durante un año. Y se utilizan polinomios a trazos y Splines para suavizar los datos y visualizar las tendencias a largo plazo en las temperaturas.

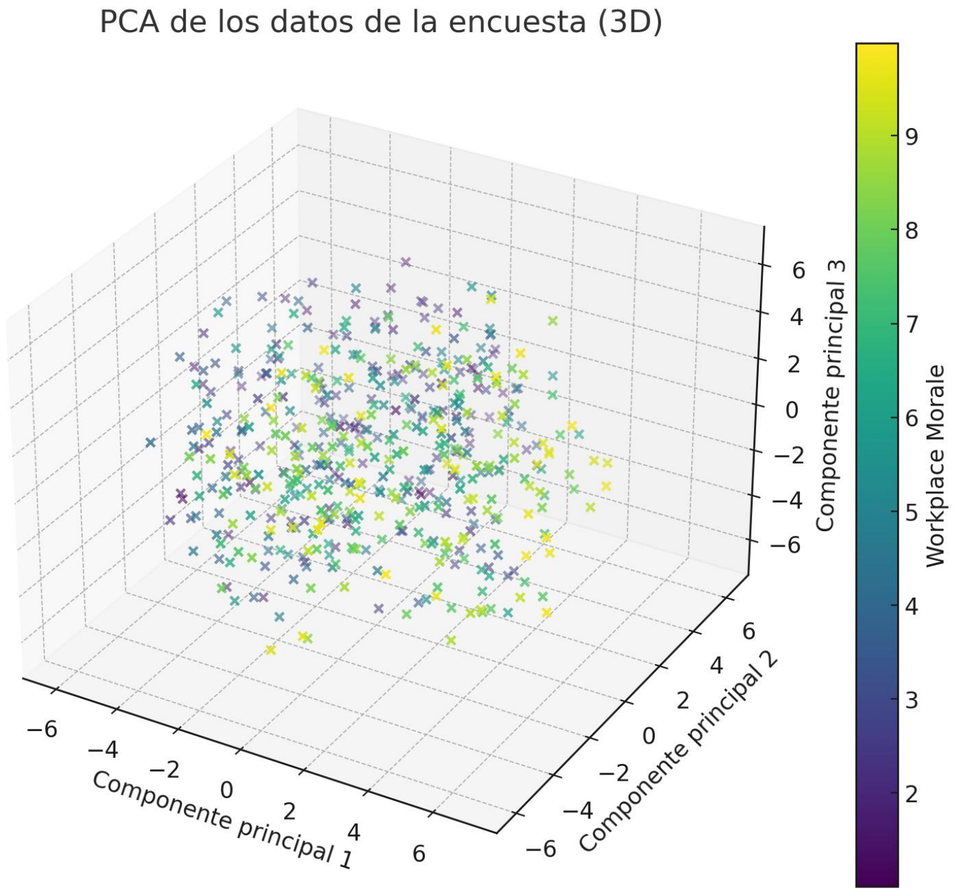
Modelos Generalizados Lineales



- Concepto de modelos generalizados lineales (GLM).
- Aplicaciones de modelos GLM: regresión logística y regresión Poisson.

Represente a los estudiantes que pasan un examen en función de las horas que estudiaron para el examen

Variables Latentes y Análisis de Datos Estructurados



- Análisis de componentes principales y el análisis de componentes independientes (ICA) como métodos para trabajar con variables latentes y datos estructurados.

Conjunto de datos de una encuesta sobre el bienestar emocional de los empleados en una empresa.

Se realiza el análisis de componentes principales (ACP) para identificar patrones subyacentes en los datos y agrupar variables relacionadas en componentes latentes

Análisis de Datos Dependientes

- Análisis de series de tiempo, definición de estacionariedad y los métodos para trabajar con datos dependientes en el tiempo
- Modelos de dependencia en datos longitudinales.

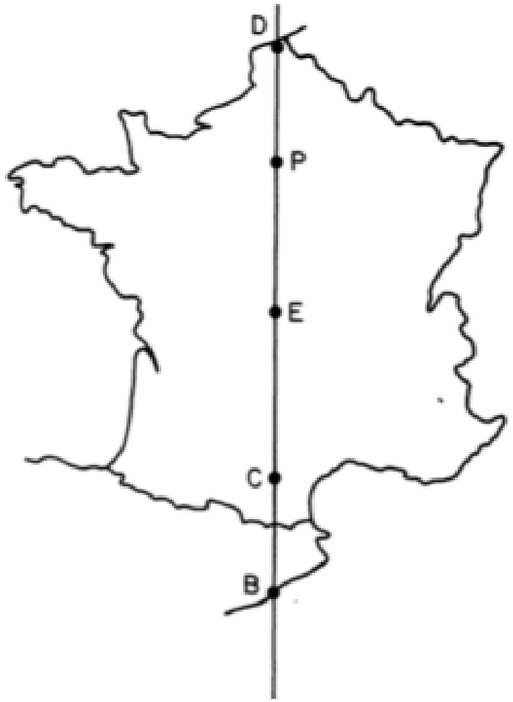
Conjunto de datos de series temporales de precios de acciones de una empresa.

Se utilizan métodos de análisis de series de tiempo para identificar patrones estacionales y tendencias en los precios y realizar predicciones a corto plazo.

ACEPTAR EL ERROR !

Con los brazos abiertos

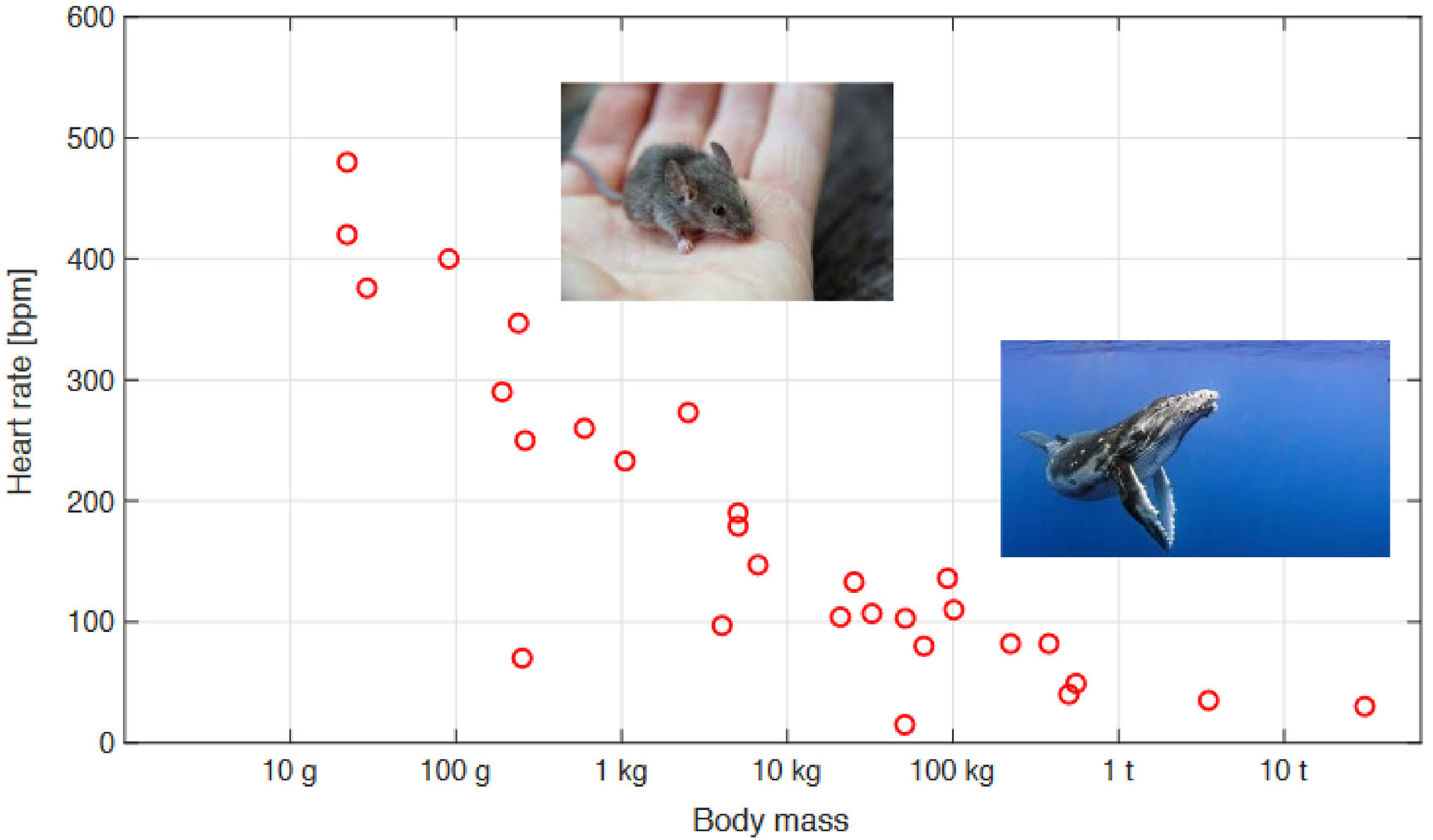
¿A qué distancia está el ecuador del polo norte?



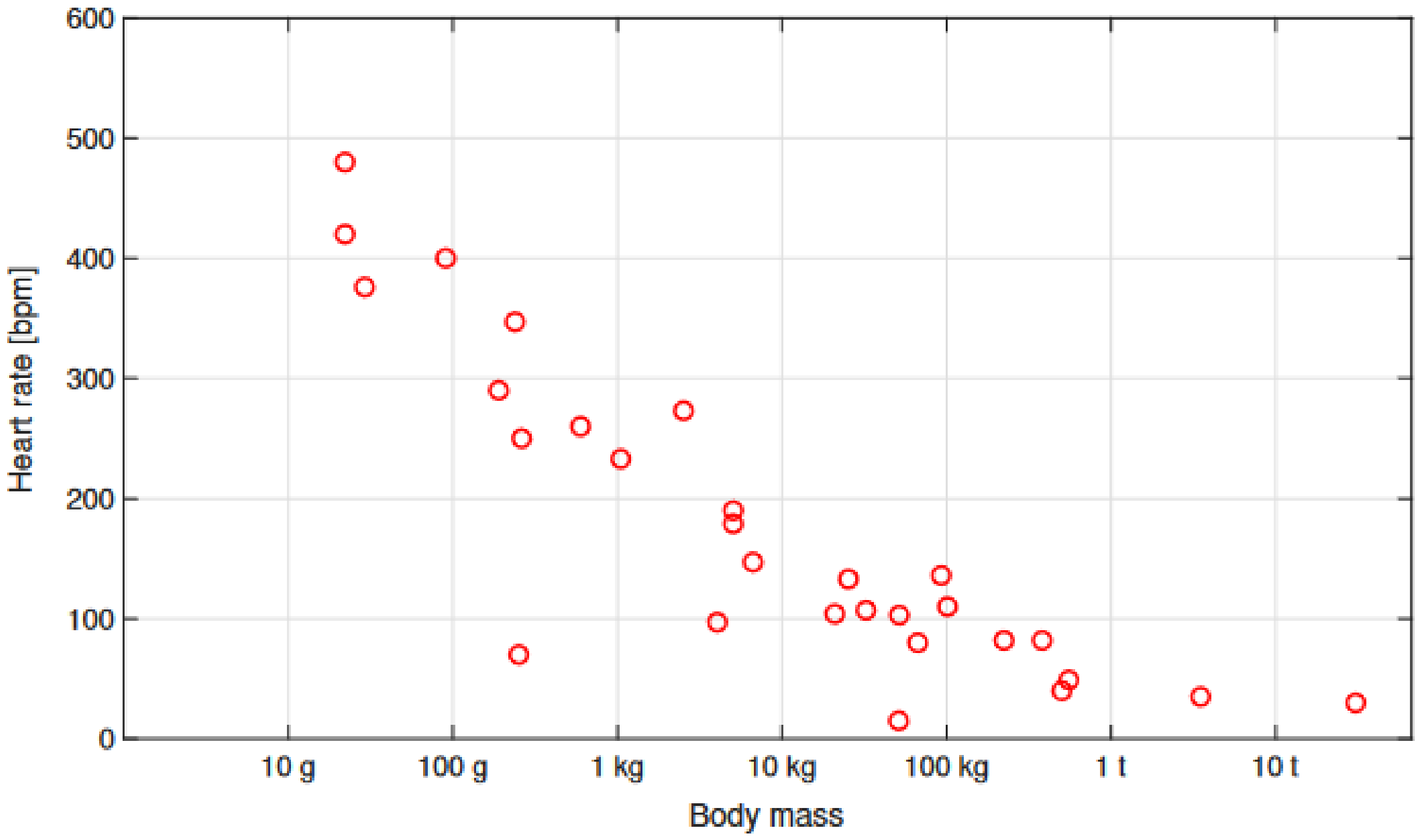
"By using this method, a sort of equilibrium is established between the **errors** which prevents the extremes from prevailing [...] [getting us closer to the] **truth**."

Adrien-Marie Legendre, 1805

Del ratón a la ballena



Del ratón a la ballena, pasando por el conejo



Los latidos de un conejo son:

(a) ≤ 100 bpm

(b) ≥ 300 bpm

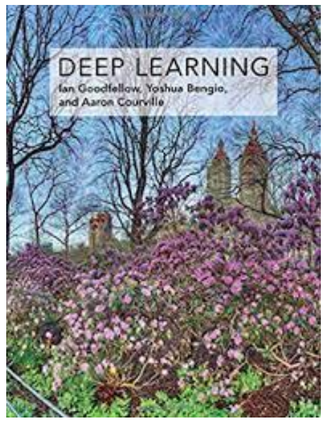
(c) ≥ 100 bpm and ≤ 300 bpm

Lo mas interesante es como hemos llegado a esta respuesta

1 Construimos la relación
2 Qué se del conejo

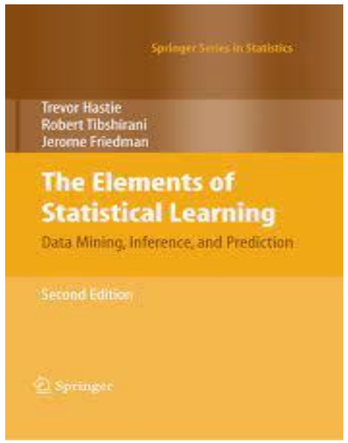
Construimos un modelo a partir de los datos
Utilizamos el modelo con el conocimiento que tenemos peso

Machine learning or statistical learning?



The ability to **acquire knowledge**, by extracting patterns from raw **data**
(Goodfellow, Bengio, Courville)

A set of tools for **modeling** and **understanding** complex **datasets**.
(James, Witten, Hastie, Tibshirani)

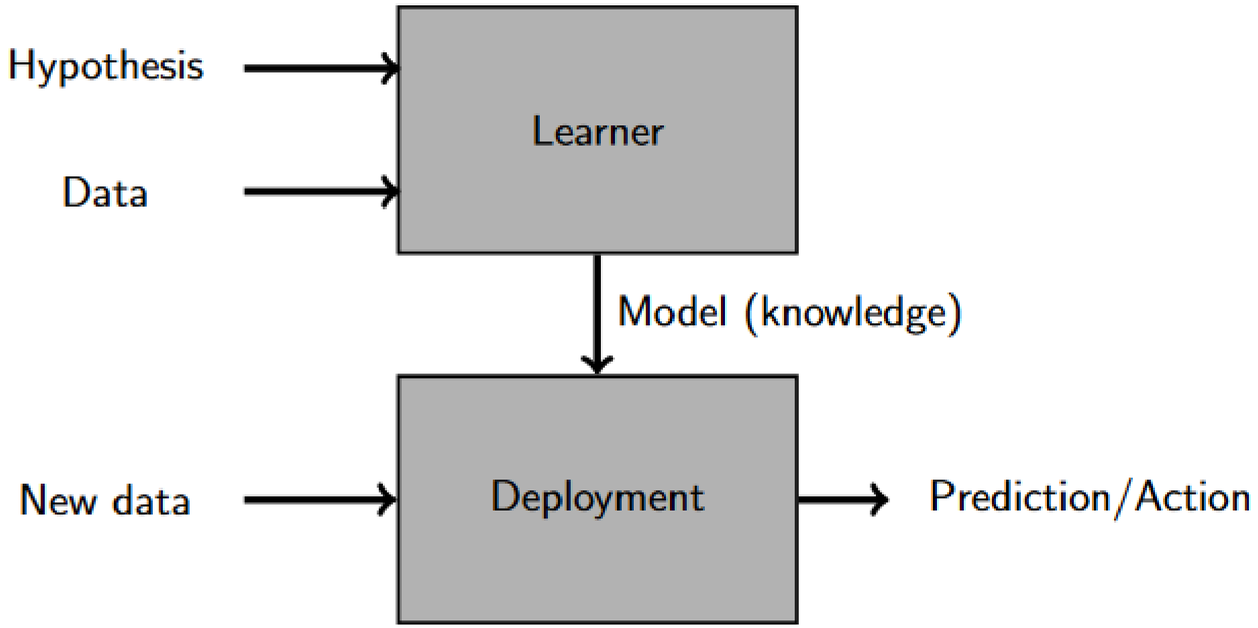


Si no hay **datos** no hay **machine learning**
El DataSet son colecciones de **elementos** descritos por un una serie de atributos.

Los modelos

Los **modelos** pueden ser construidos, vendidos e implementados para entregar valor.

Durante la vida de un modelo, podemos distinguir dos etapas.
1. Etapa de aprendizaje: Se construye el modelo.
2. Etapa de implementación: Se utiliza el modelo.



Extraer conocimiento

Y cuál modelo?

Nos interesa encontrar el **mejor modelo**. Por lo tanto, necesitamos una noción de **calidad de modelo**.
Realmente, nuestro objetivo es construir modelos que funcionen **bien durante la implementación**, es decir, cuando se presentan con nuevos datos.

Una metodología básica de ML siempre incluye dos tareas separadas:

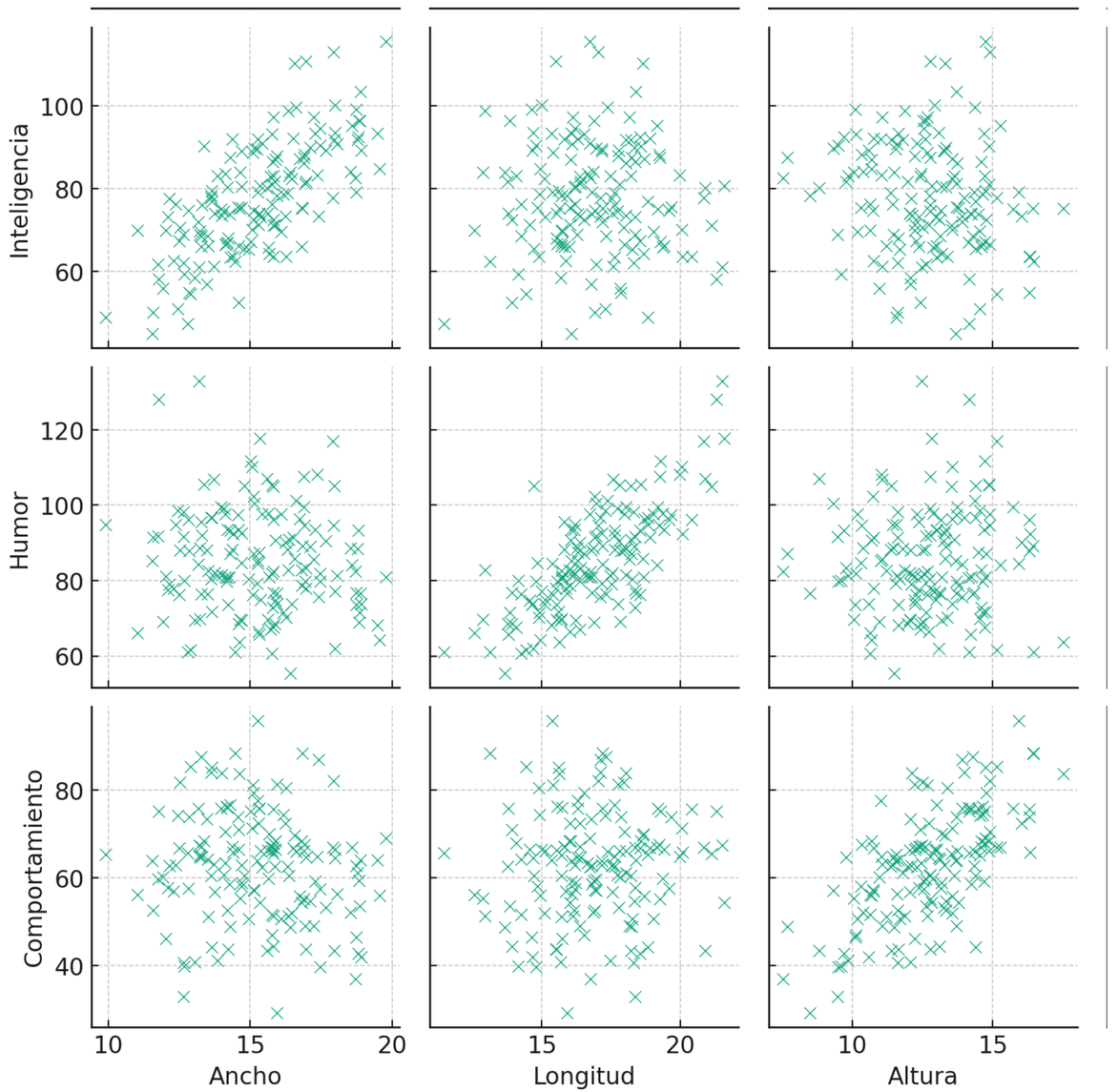
Training: se crea un modelo utilizando datos y una métrica de calidad. También decimos **que ajustamos un modelo** a **Dataset**

Testing: se evalúa el rendimiento del modelo durante la implementación utilizando datos nuevos e **inéditos**.

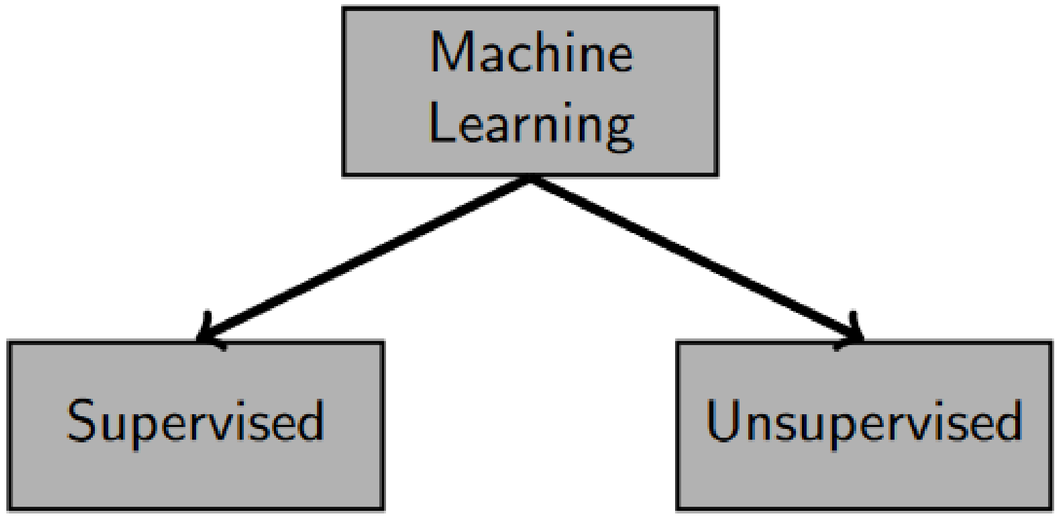
Como evaluamos

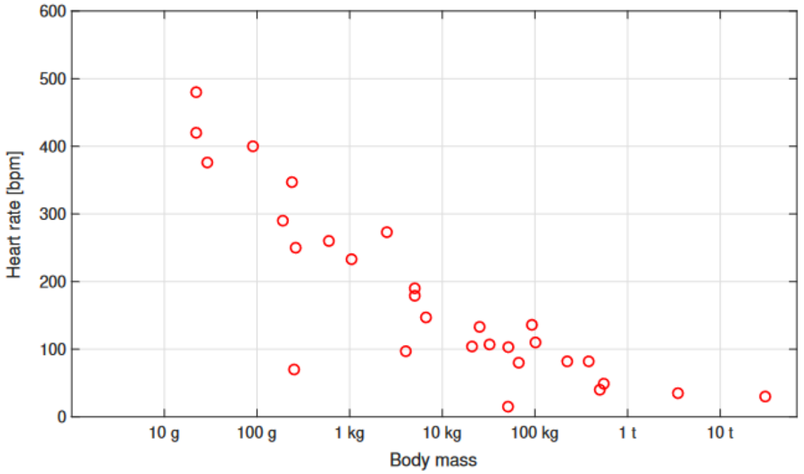
La craniometría siglo l siglo XIX

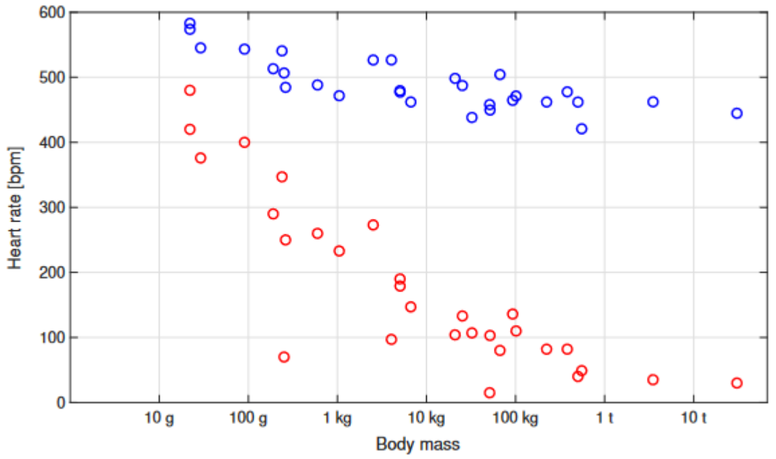
Los craniometristas medían las dimensiones del cráneo humano, como el ancho, la longitud y la altura, y creían que podían establecer correlaciones entre estas mediciones y ciertos aspectos de la personalidad y el comportamiento.



Qué tipo de problemas se pueden formular?





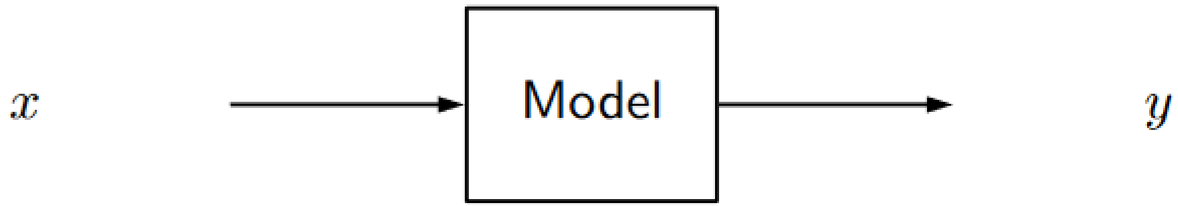


Aprendizaje supervisado

En aprendizaje supervisado, se nos presenta un **nuevo elemento** (un conejo) y **no conocemos el valor** de una de sus atributos (su ritmo cardíaco).

Nuestro objetivo es **estimar** (adivinar) **el valor faltante** a partir de una **colección de elementos conocidos** (peso y ritmo cardíaco de otros animales).

El desafío es construir un modelo que mapee un atributos ***x*** conocido como el **predictor** a otra característica ***y*** que llamamos la **etiqueta**, utilizando un conjunto de datos de ejemplos etiquetado

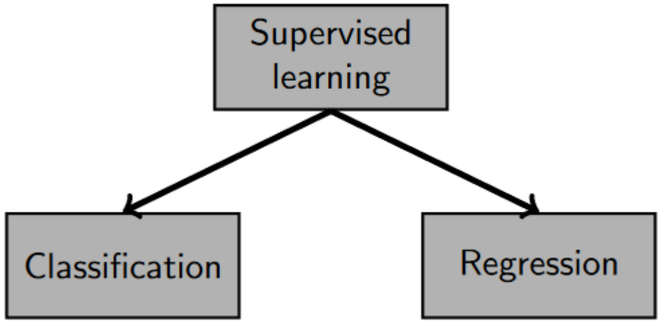


Aprendizaje supervisado: Clasificación y regresión

El aprendizaje supervisado se divide en dos categorías dependiendo del tipo de etiqueta:

- **Clasificación**: La etiqueta es una variable **discreta**.
  En un detector de spam, **0** podría significar que el correo electrónico es spam y **1** que no lo es.
  Predicción que emoción hay detrás del tuit, positiva, neutral, negativa

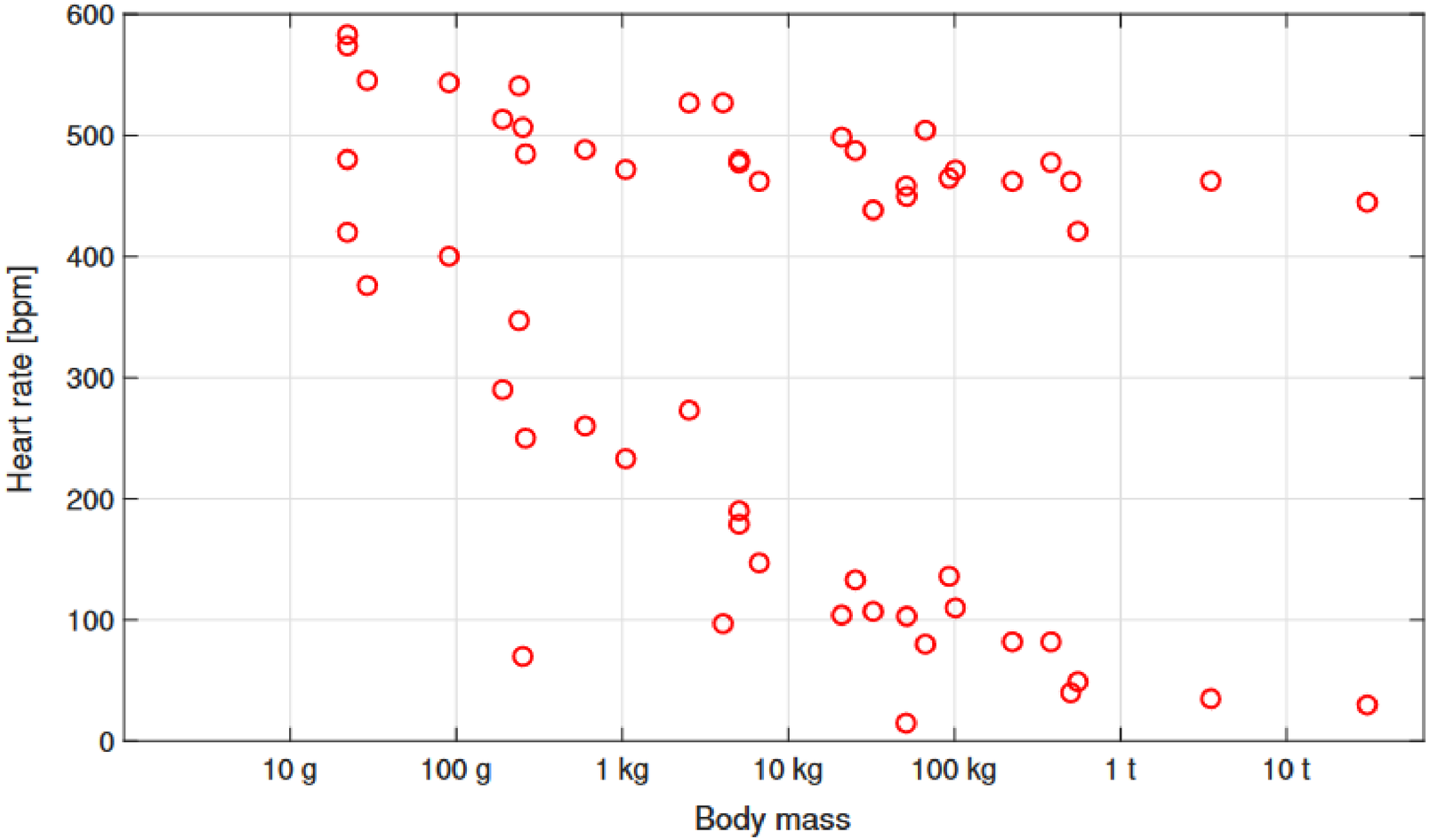
- **Regresión**: La etiqueta es una variable **continua**.
  El ritmo cardíaco de un animal es una etiqueta continua



Aprendizaje No supervisado:
Ritmo cardíaco en el zoológico galáctico

Aprendizaje No supervisado:
Ritmo cardíaco en el zoológico galáctico

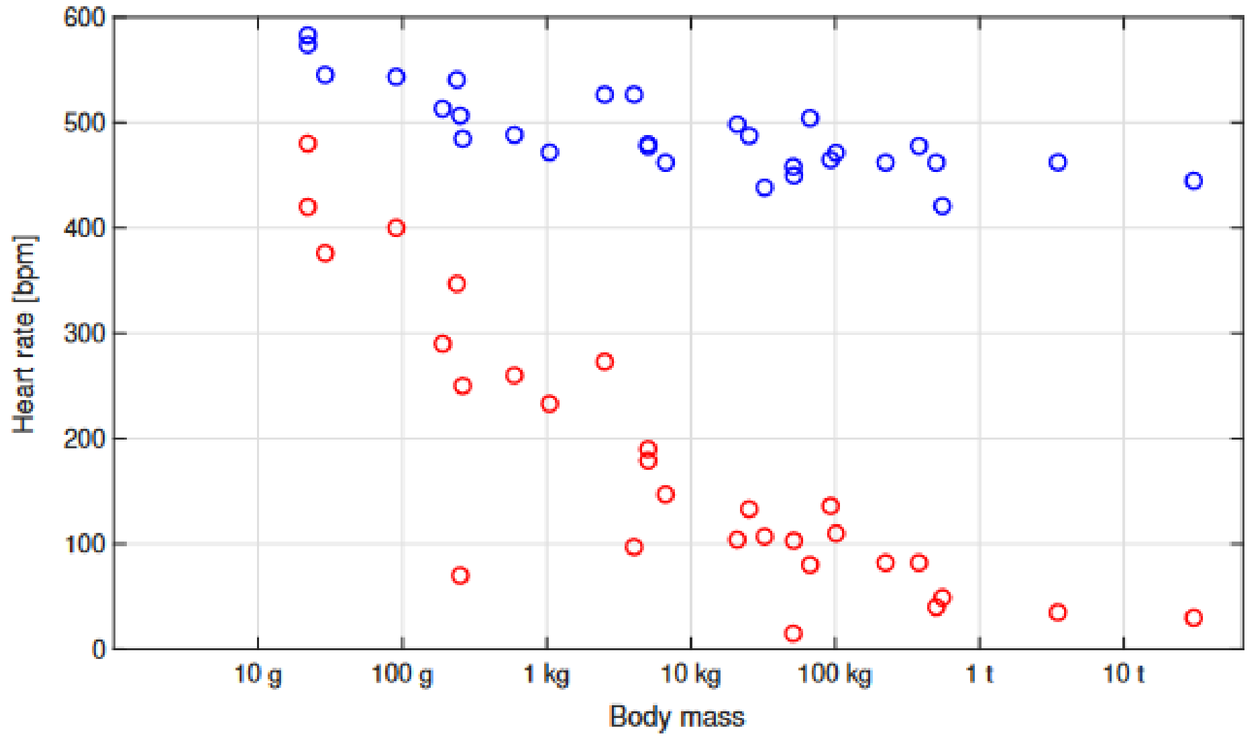
Qué podemos concluir de esta distribución?



Aprendizaje No supervisado

En el aprendizaje no supervisado, nuestro objetivo es **encontrar la estructura de Dataset, cómo se distribuyen los datos** . Entre otros usos, esto puede ser útil para 1) la comprensión de los datos, 2) identificar anomalías, 3)comprimir nuestros datos y reducir el tiempo de procesamiento.

No queremos predecir

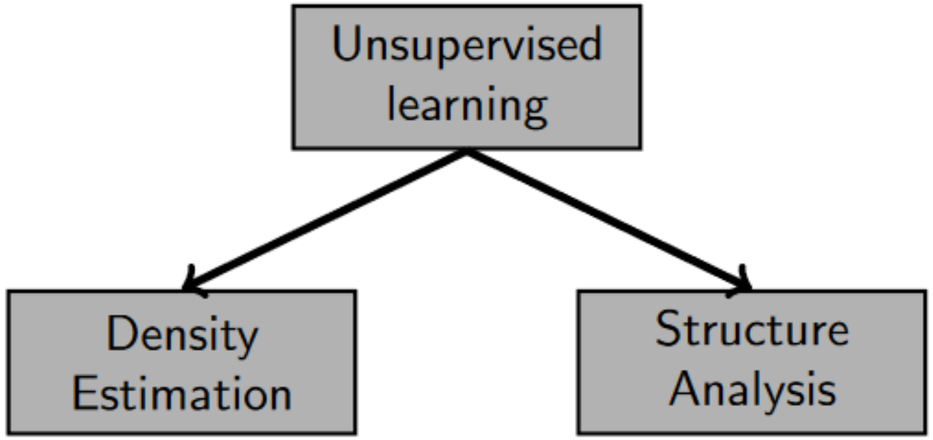


Aprendizaje No supervisado

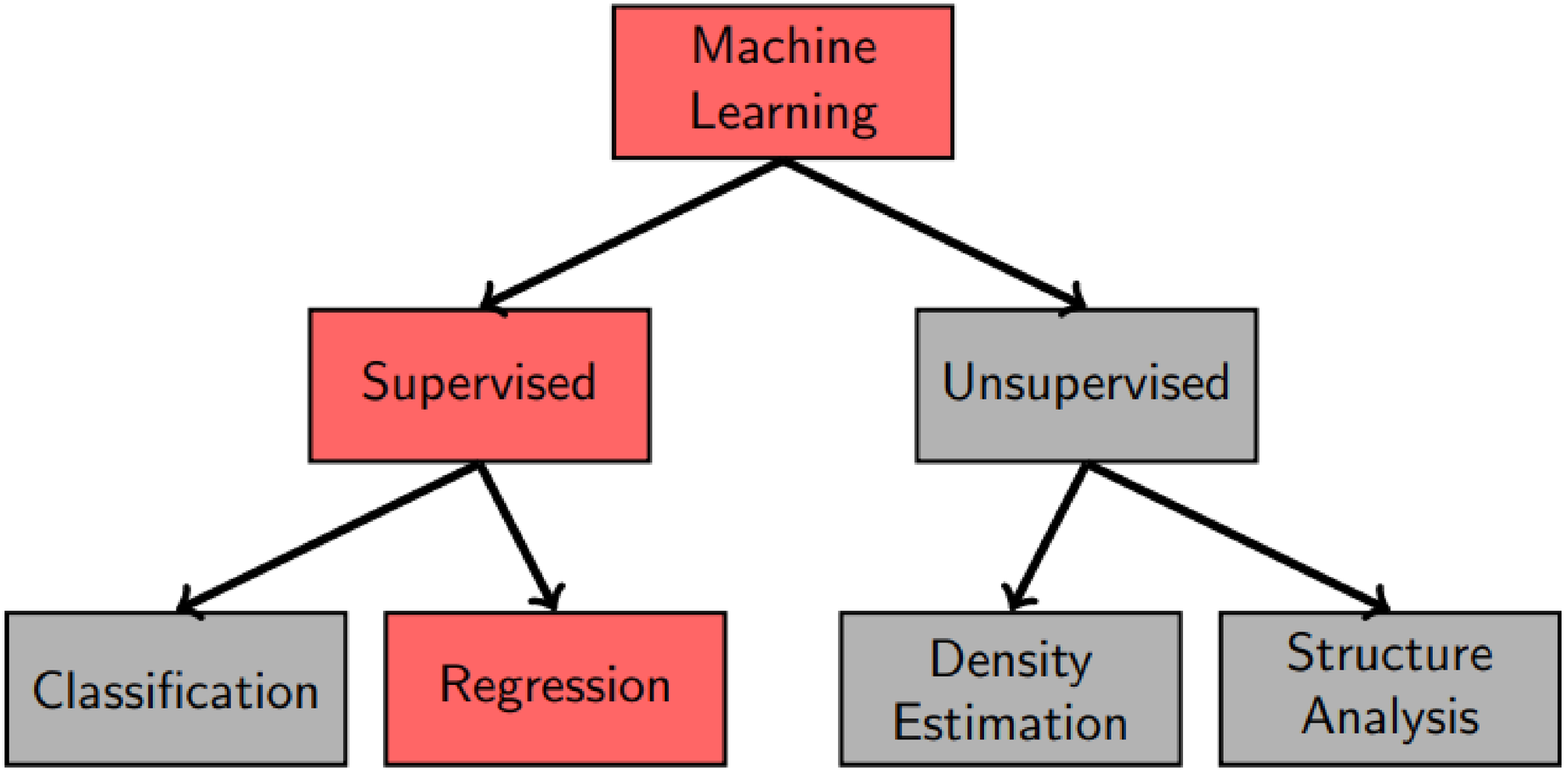
Análisis de la Densidad: La técnica proporcionan modelos para la distribución de muestras en el espacio de atributos. Probabilidad de encontrar una serie de puntos en una región del espacio,

Análisis de la estructura el dataset también puede estudiarse mediante el análisis de estructuras, que incluye:

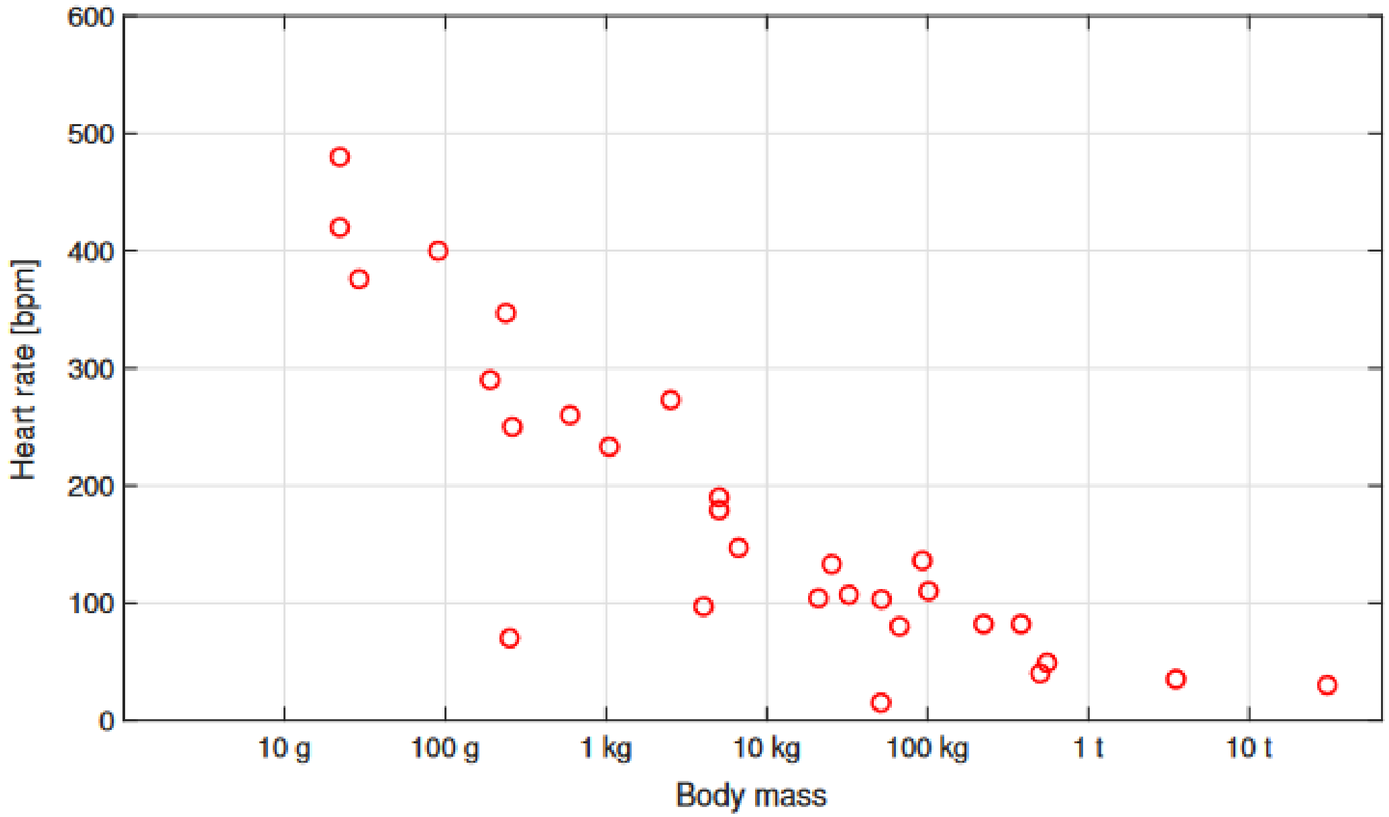
Análisis de cluster: se centra en grupos de puntos de datos.
Análisis de componentes: identifica las direcciones de interés.



Taxonomía


Machine Learning
Supervised
Unsupervised
Classification
Regression
Density Estimation
Structure Analysis

El extraño caso del gusano plano



El ritmo cardiaca
gusanos planos (flatworm) que
pesa menos de 10g

a) No podemos adivinar con este dataset
b) > 300 bpm
c) Ninguna de las anteriores

Formulación de problemas de regresión

La regresión es un problema **supervisado**: Asumimos que el valor de uno de los atributos (etiqueta) puede ser predicho a partir del valor de los atributos restantes (predictores).

La etiqueta es **una variable continua**.

Nuestro trabajo entonces es **encontrar el mejor modelo que asocie** una única etiqueta a un conjunto dado de predictores.

Predictores y etiquetas

	Age	Salary
S_1	18	12000
S_2	37	68000
S_3	66	80000
S_4	25	45000
S_5	26	30000
...	...	...

(a) La Edad es el predictor, el Salario es la etiqueta.
(b) El Salario es el predictor, la Edad es la etiqueta.
(c) Ambas opciones pueden ser consideradas.

Association and causation

Los modelos de predicción a veces se interpretan a través de una lente causal: el **predictor** es la **causa**, la **etiqueta** su **efecto**. Esto **no es correcto**

Nuestra capacidad para construir predictores se debe a la **asociación** entre los **atributos**, más que a la causalidad. Dos atributos en un conjunto de datos parecen estar asociados:

Si uno causa al otro (directa o indirectamente).
Cuando ambos tienen una causa común.
Debido al muestreo.

Construimos modelos peor no son casuales

Notación Matemática

Dataset:

N es el numero de muestras, ***i*** identifica cada muestra

x_i es el **predictor** de la miestra ***i***

y_i es la **etiqueta realera** continua de la muestra ***i***

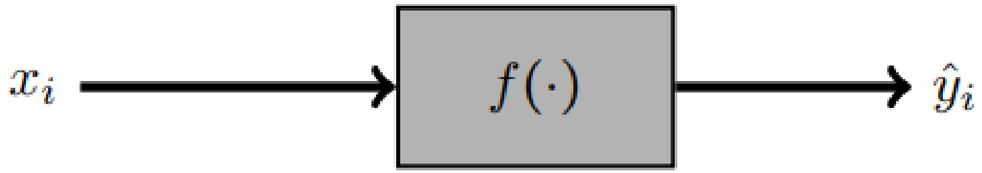
El **dataset** es $\{(x_i, y_i) : 1 \leq i \leq N\}$ y (x_i, y_i) es la muestra ***i***

Model:

$f(\cdot)$ denota el modelo

$\hat{y}_i = f(x_i)$ es la etiqueta predecida de la muestra ***i***

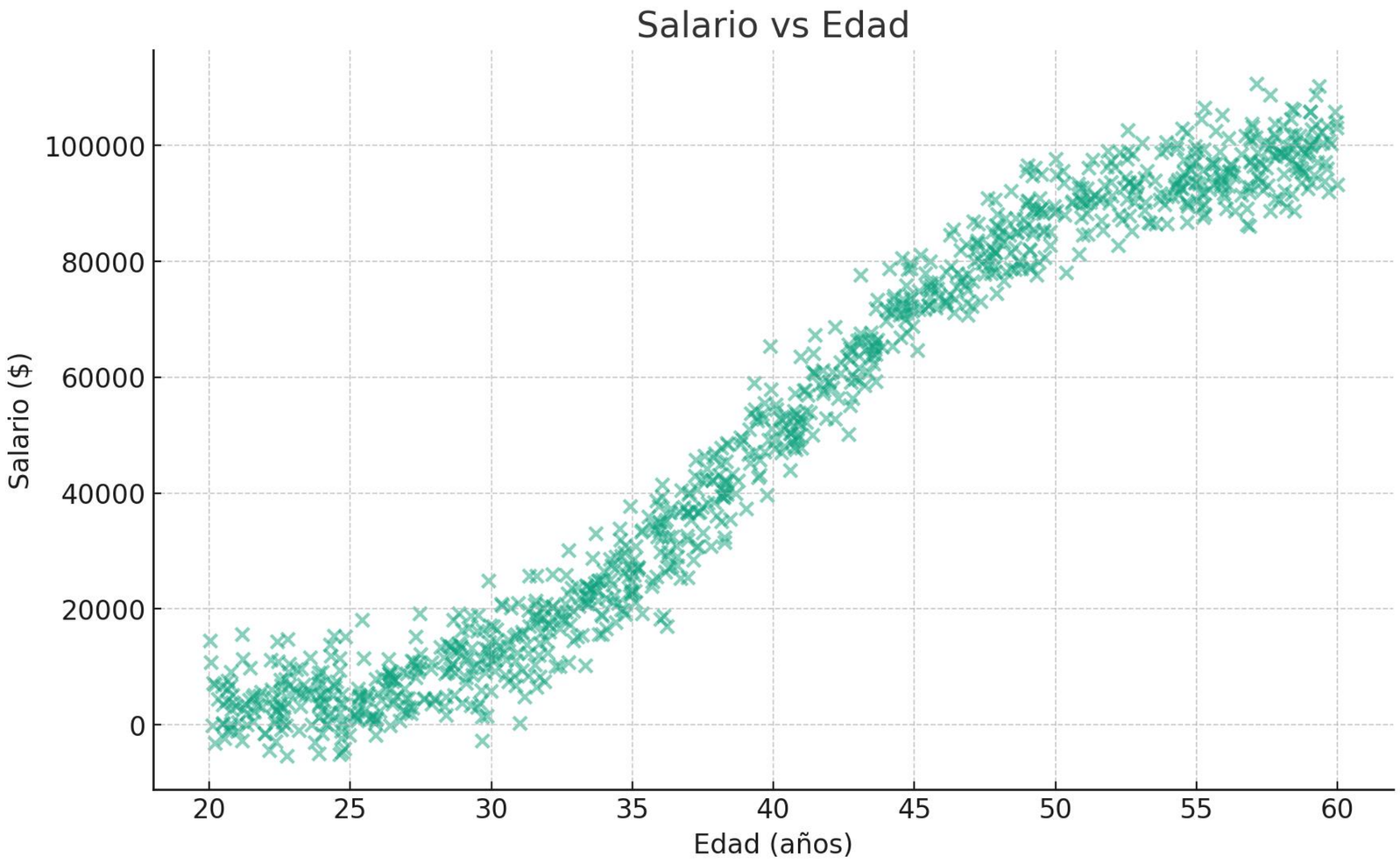
$e_i = y_i - \hat{y}_i$ es el erro de la predicción de la miestra ***i***



(Ten en cuenta que estamos considerando un predictor aquí, esta notación se extenderá a múltiples predictores cuando se discutan modelos multivariados)

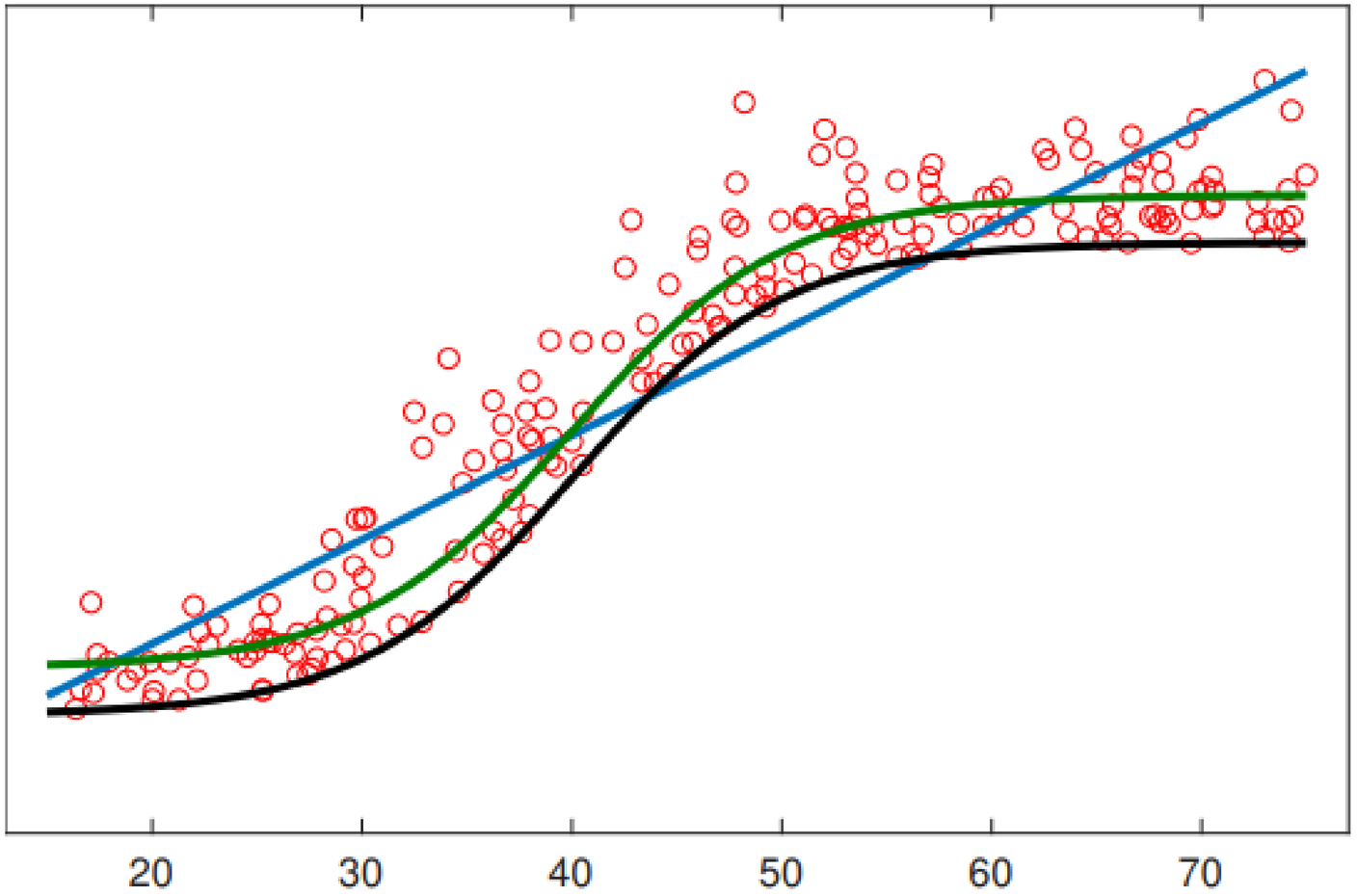
Regresión

¿Cuál línea es la mejor representación de la edad con respecto al salario?



Candidato a la solución

¿Cuál línea es la mejor representación de la edad con respecto al salario?



Cuál es el mejor modelo ?

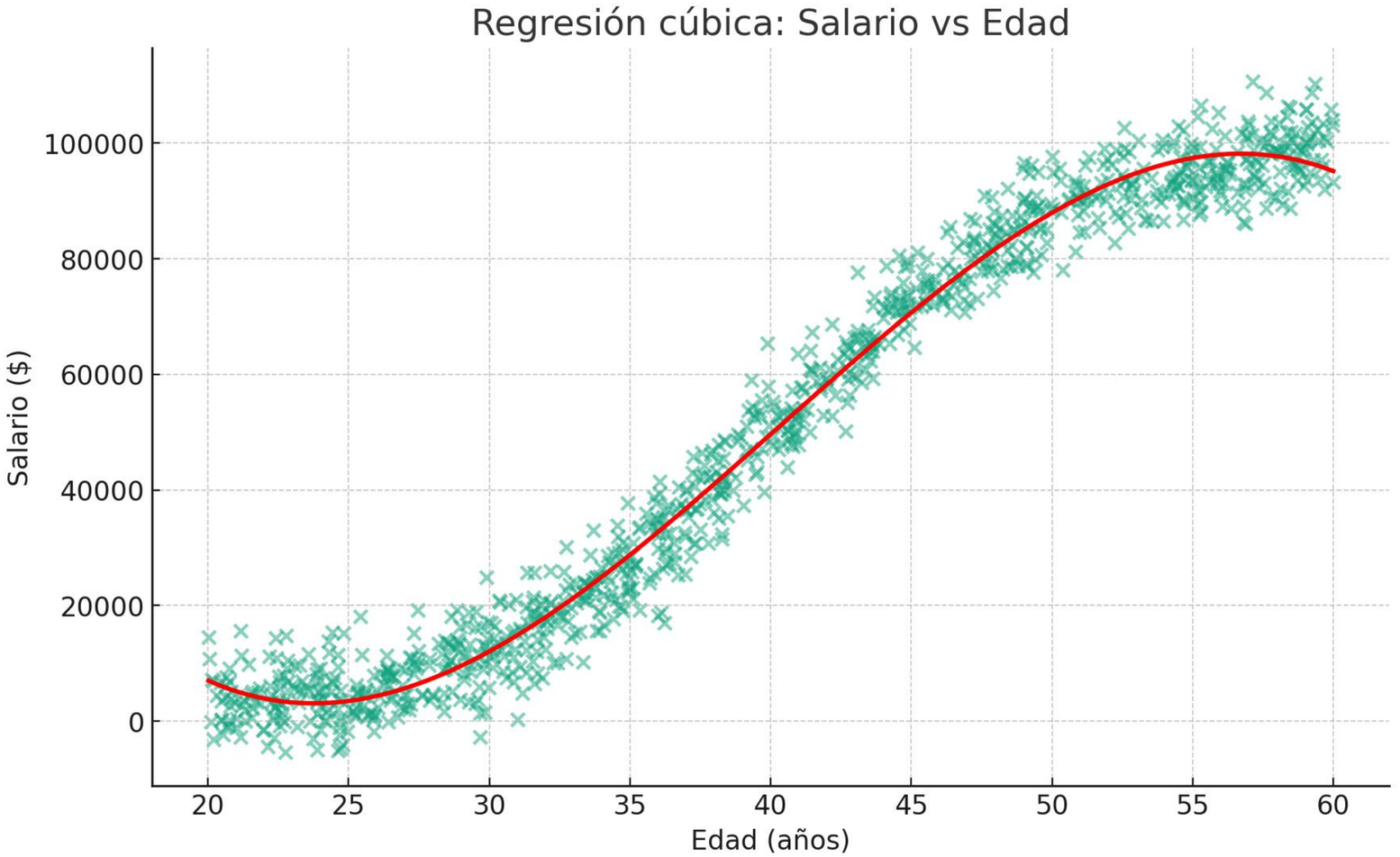
Para que podamos encontrar el **mejor modelo**, necesitamos una noción de **calidad del modelo**.
Una métrica de calidad popular en problemas de regresión es el **error cuadrático medio (MSE),** que corresponde al error cuadrado esperado de la predicción de un modelo durante su despliegue.

Si se nos da un conjunto de datos que consta de **N** muestras y un modelo ***f(·)***, podemos estimar su **MSE** de la siguiente manera:

$$
\begin{aligned}
E_{MSE} &= \frac{1}{N}\sum_{i=1}^{N} e_i^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} (y_i - f(x_i))^2
\end{aligned}
$$

Mejor modelo según noción de calidad

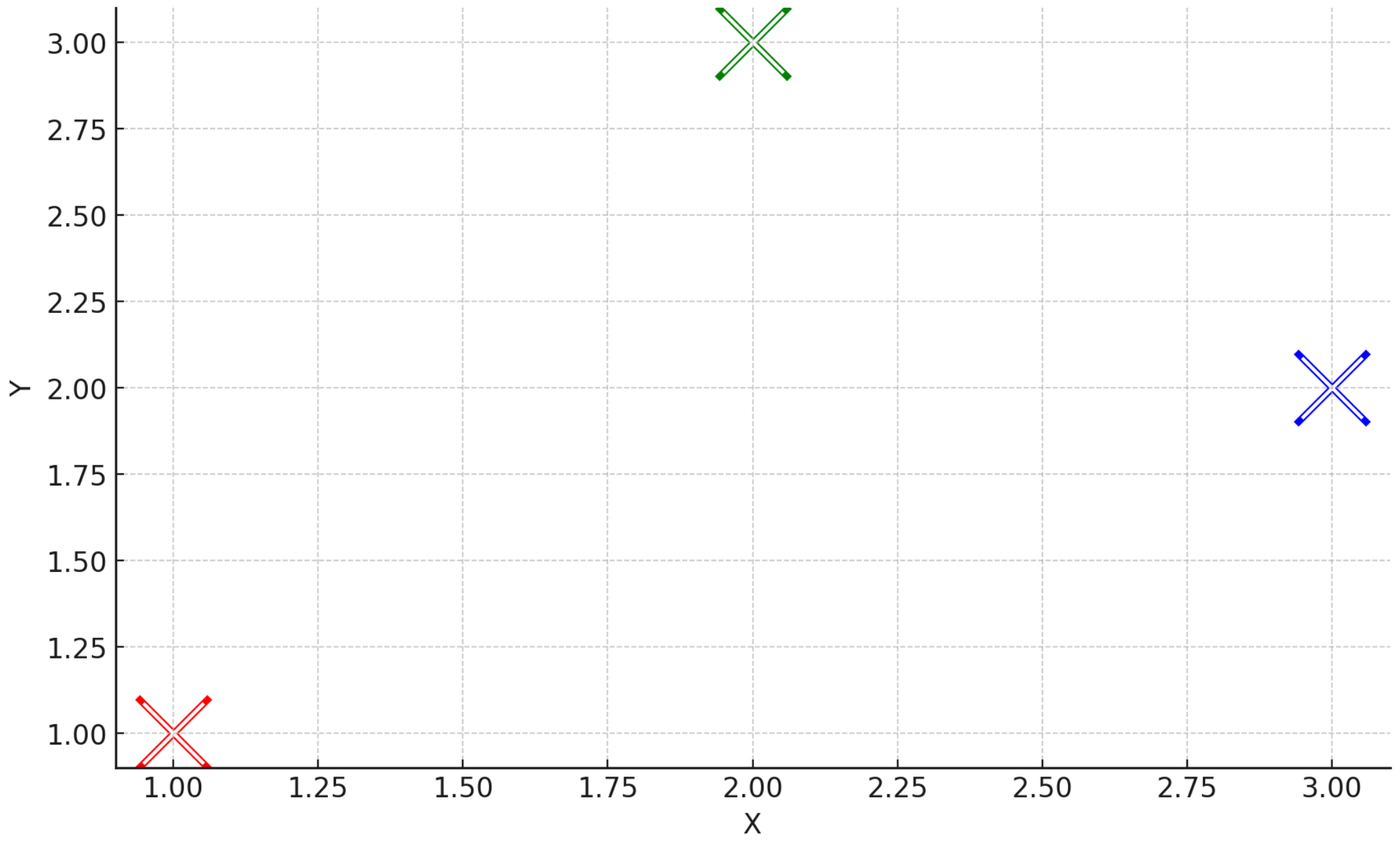
Si tenemos en cuenta la noción de calidad MSE este seria el mejor modelo



EJEMPLO MSE

$$E_{MSE} = \frac{1}{N}\sum_{i=1}^{N} e_i^2$$

X	Y
1	1
2	3
3	2



MSE para $f =$ X

MSE para un f que una todos los puntos

¿Un modelo sin errores?

Dado un dataset, ¿es posible encontrar un modelo tal que $\hat{y}_i = y_i$ para cada instancia ***i*** en el dataset ?, es decir, un modelo cuyo error es cero, EMSE = 0?

(a) Nunca, siempre habrá un error no nulo

(b) No esta garantizado pero podría ser posible para algunos conjuntos de datos

(c) Siempre, siempre habrá un modelo lo suficientemente complejo que lo logre

La naturaleza del Error

Cuando consideramos un problema de regresión, necesitamos ser conscientes de que:

- Los **predictores** elegidos pueden **no incluir todos los factores** que determinan la etiqueta.
- El **modelo elegido puede no ser capaz de representar con precisión** la verdadera relación entre la respuesta y el predictor (el patrón).
- Pueden estar presentes mecanismos **aleatorios** (ruido).

Matemáticamente, representamos esta discrepancia como

$$\begin{aligned} y &= \hat{y} + e \\ &= f(x) + e \end{aligned}$$

Siempre habrá alguna discrepancia (error e) entre la verdadera etiqueta y y nuestra predicción de modelo f(x).

¡Acepta el error!

La regresión como un problema de optimización

Dado un conjunto de datos {(xi, yi) : 1 ≤ i ≤ N}, cada modelo candidato f tiene su propio EMSE. Nuestro objetivo es encontrar **el modelo con el EMSE** más bajo:

$$f_{best}(x) = \arg\min_{f} \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2$$

La pregunta es, ¿cómo encontramos dicho modelo?
Encontrar tal modelo es **un problema de optimización.**

Obsérvese que estamos buscando el modelo que minimice el EMSE del dataset, sin embargo, este modelo puede que no sea la solución de **error cuadrático medio mínimo** (MMSE),

¡es decir, el mejor modelo durante la implementación!