Practical Data Science:

Intro to Machine Learning & Linear Regression

D. Sierra-Porta
November 2021

A continuación...

- Machine Learning
  - Tipos de problemas que resuelve Machine Learning
  - Pasos para resolver problemas con Machine Learning
- Importancia de Machine Learning
- Casos de estudio

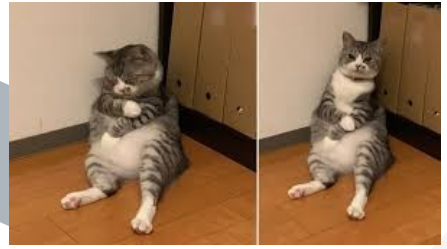
Qué es un gato?




llamado popularmente gato, y de forma coloquial minino, michino, michi, micho, mizo, miz, morroño o morrongo, entre otros, es un mamífero carnívoro de la familia Felidae. Es una subespecie domesticada por la convivencia con el ser humano.

Diversidad


Deformación

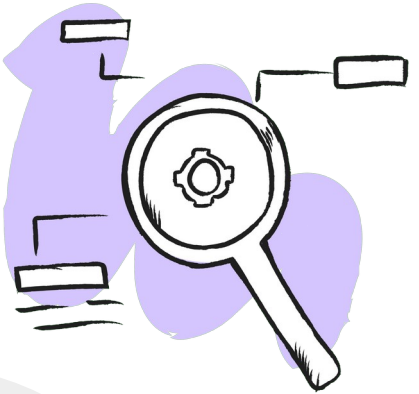



Oclusion


Condiciones de luz?

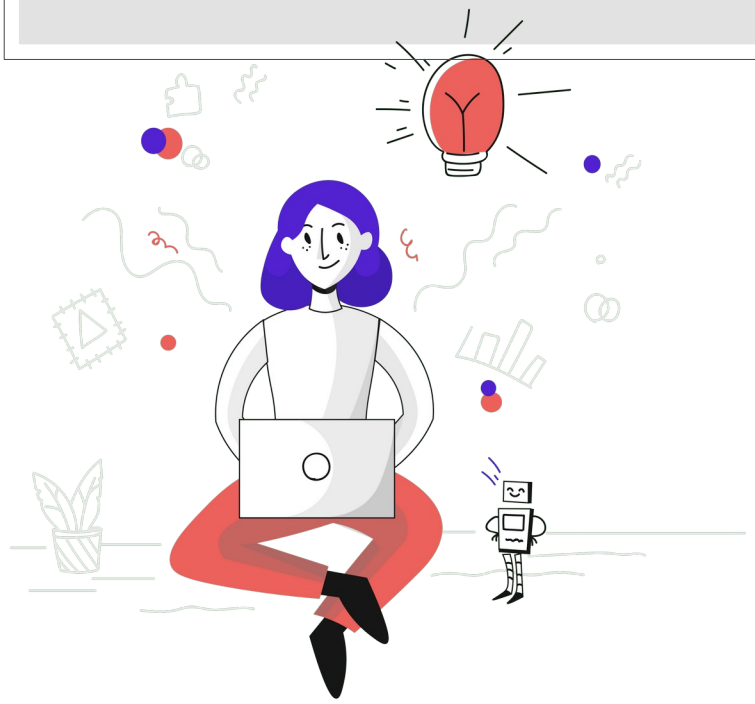


¿Qué es Machine Learning?



"field of study that gives computers the ability to learn without being explicitly programmed"

Arthur Samuel, 1959



"A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**." ...

Tom Mitchell, 1997

Using data for **answering questions**

Training Predicting

- Datos disponibles en cualquier lado
- Economia del almacenamiento → **for free**
- Hardware más poderoso y rápido que hace décadas

Datos

- Prácticamente todos tenemos un computador completamente llenos de sensores: GPS, cámaras, micrófonos...
- Estamos permanentemente conectados a la internet

Dispositivos

- Cloud computing:
  - Almacenamiento en nuve
  - Infraestructura como un servicio
- Aplicaciones de usuario:
  - Youtube, Twitter, Gmail, Facebook, Instagram,...

Servicios

Tipos de problemas de Machine Learning

Supervisados

No supervisados

Refuerzo

Tipos de problemas de Machine Learning

Supervisados

Aprendiendo a través de ejemplos de los cuales conocemos el resultado deseado (lo que queremos predecir)

- ¿Es un perro o un gato?

No supervisados

- ¿Un mail es spam o no lo es?

Refuerzo

- Predecir el valor de mercado de una casa dependiendo de su tamaño, ubicación, número de cuartos, vecindario, etc...

Tipos de problemas de Machine Learning

Supervisados

No supervisados

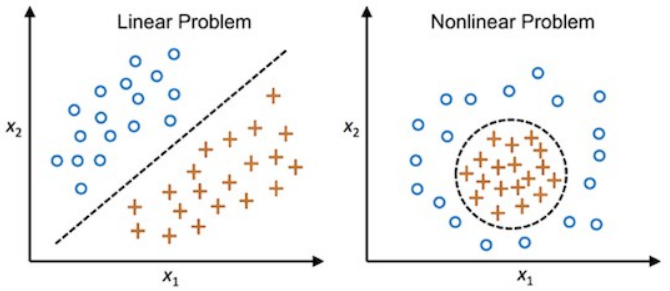
Refuerzo

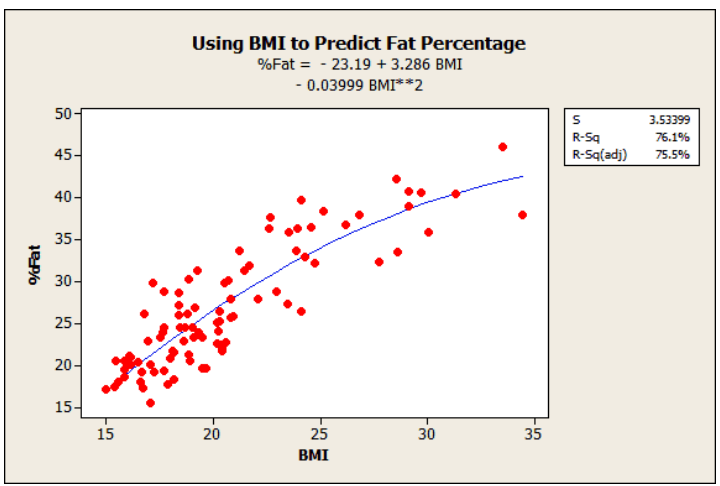
Clasificación

El resultado es una variable Discreta (ej. perro, gato)

Regresión

El resultado es una variable continua (ej. temperatura, tamaño, precio, voltaje, etc...)





Tipos de problemas de Machine Learning

Supervisados

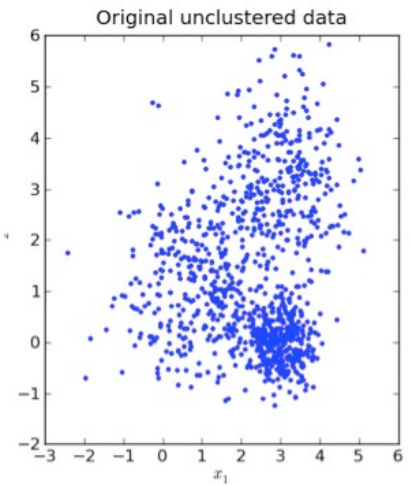
No supervisados

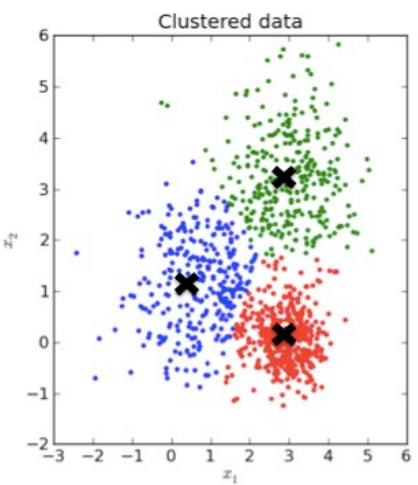
Refuerzo

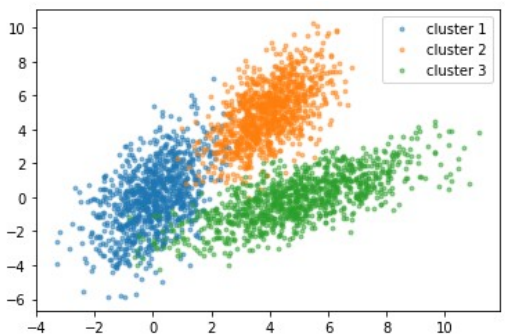
No existe un resultado deseado. Se aprende de todo acerca de los datos. Relaciones latentes.

- De un conjunto de fotos, ordenar grupos de 20 fotos.
- Encontrar anomalías en los patrones de consumo de tarjetas de crédito.

Muy útil para encontrar estructura en los datos (agrupamiento de datos), correlaciones escondidas, reducir dimensionalidad, etc...







Tipos de problemas de Machine Learning

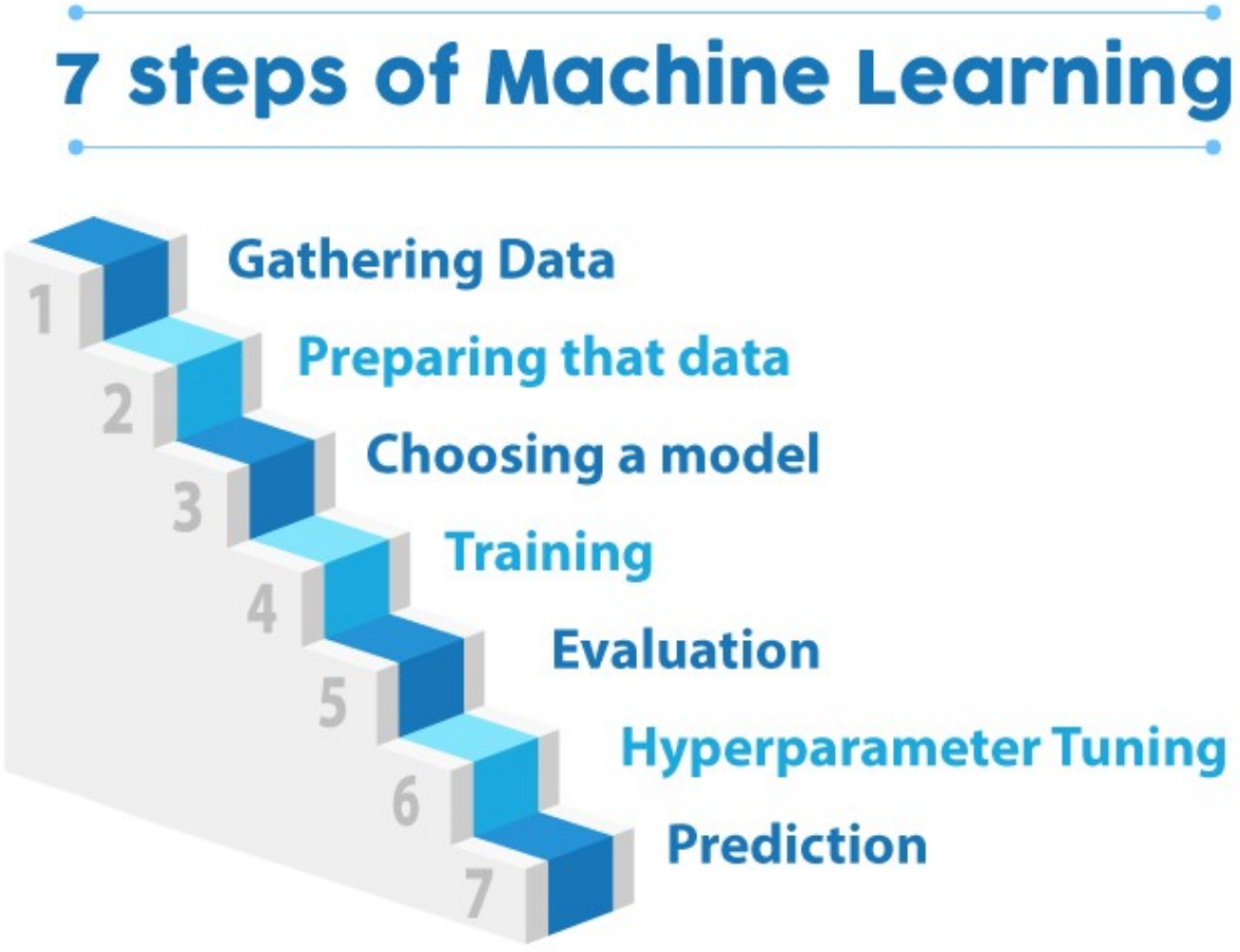
Supervisados

No supervisados

Refuerzo

- **Primero**: se interactua con el ambiente y se aprende de las condiciones y situaciones observando relaciones entre los datos.
- **Segundo**: Se establece un feedback de acuerdo a un reforzamiento positivo o uno negativo

Pasos para resolver un problema de Machine Learning



Manos a la obra...

GASTRO

Abalón, el molusco más caro del mundo que se cría en A Coruña

- ¿Una cena romántica? Le proponemos ostras para herejes
- Cómo cocer percebes: trucos para cocinar el marisco que comerá esta Navidad
- Gamba Natural, una granja pionera que cría langostinos en Medina del Campo
- Cómo preparar un buen centollo: así se cuece el changurro

Muestra de varios abalón en su concha. Al tratarse de un animal marino muy musculado, retirar la carne es un proceso muy delicado | Luis de las Alas

Javier Caballero

Actualizado: 02/01/2017 10:03 horas

https://www.expansion.com/fueradeserie/gastro/2016/12/29/585cfec7ca474125398b4580.html

https://www.unisinucartagena.edu.co/curso-potencial-de-moluscos-en-colombia-para-produccion-de-perlas/#1586388071970-9c842761-440e

ETAPAS PARA EL CULTIVO DE BIVALVOS MARINOS (PECTÍNIDOS Y OSTRAS) EN SISTEMA SUSPENDIDO EN EL CARIBE COLOMBIANO

SHIA WANE SUKUAIPA SÜPÜLA EPIJAA O APUNAJAA WANE SHE>E O SE>E PALAA (WARUTTA) TÜÜ SUKUAIPA YAA SHIROKU PALAA PALAMUNKA SULU>U COLOMPIA

JAVIER GÓMEZ-LEÓN
OLGA L. LARA QUINTERO
CAMILA ROMERO CHICA

Instituto de Investigaciones Marinas y Costeras
"José Benito Vives De Andréis" – INVEMAR , Santa Marta, Colombia
Vinculado al Ministerio de Ambiente, Vivienda y Desarrollo Territorial
Cerro Punta Betín, Santa Marta, DTCH
PBX (+57) (+5) 4380808 • Fax (+57) (+5) 4233280 • A. A. 1016 • **www.invemar.org.co**

Cítese como:
J. Gómez-León, O. Lara y C. Romero., 2009. Etapas para el cultivo de bivalvos marinos (pectínidos y ostras) en sistema suspendido en el Caribe colombiano. Serie de Publicaciones Generales Nº 25. Santa Marta, 36 Pág.

ISBN: 978-958-8448-06-0

Palabras Clave: Pectínidos, ostras, cultivo suspendido, captación, crecimiento, selección, supervivencia, cosecha

Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

A New Method of Measuring the Age of Abalone Based on Data Visualization Analysis

To cite this article: Runze Guo *et al* 2021 *J. Phys.: Conf. Ser.* **1744** 042181

Abstract. This project uses a new way to count the abalone age, which use abalone's physical characteristics to predict by multiple linear regression. After the model is trained, when we catch a new abalone, we can use a computer to replace the labor to a certain extent, saving costs to the enterprise. Results are given in a visualization of the data.

Keywords: Abalone, Multi-Linear Regression, Data Visualization

Extending and benchmarking Cascade-Correlation: extensions to the Cascade-Correlation architecture and benchmarking of feed-forward supervised artificial neural …

SG Waugh - 1995 - eprints.utas.edu.au

This thesis is divided into two parts: the first examines various extensions to Cascade-Correlation, and the second examines the benchmarking of feed-forward supervised artificial neural networks, including back-propagation and Cascade-Correlation. The first …

☆ 99 Citado por 81 Artículos relacionados Importar al BibTeX »

Abalone Data Set

Download: Data Folder, Data Set Description

Abstract: Predict the age of abalone from physical measurements

Data Set Characteristics:	Multivariate	**Number of Instances:**	4177	**Area:**	Life
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	8	**Date Donated**	1995-12-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	1166512

Source:

Data comes from an original (non-machine-learning) study:
Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994)
"The Population Biology of Abalone (_Haliotis_ species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait",
Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288)

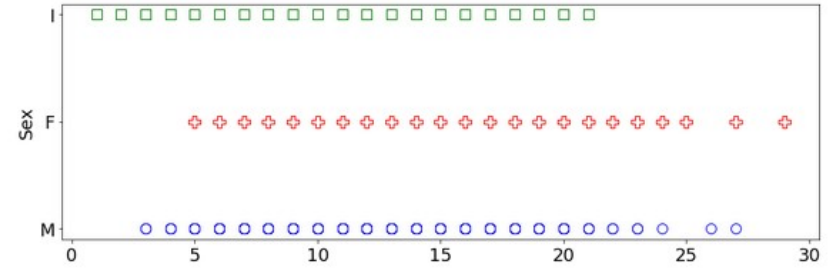
Original Owners of Database:

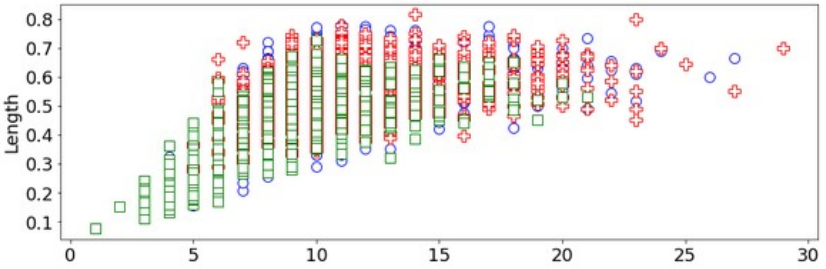
Marine Resources Division
Marine Research Laboratories - Taroona
Department of Primary Industry and Fisheries, Tasmania
GPO Box 619F, Hobart, Tasmania 7001, Australia
(contact: Warwick Nash +61 02 277277, wnash '@' dpi.tas.gov.au)

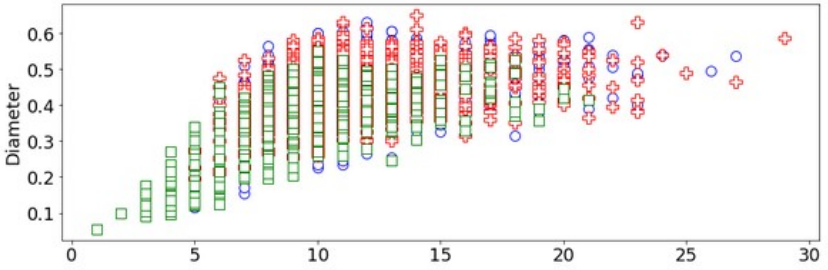
Donor of Database:

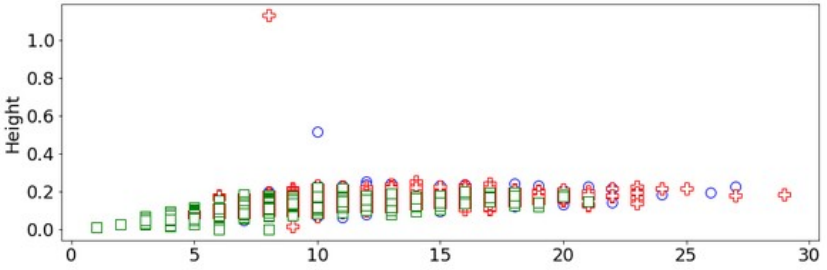
Sam Waugh (Sam.Waugh '@' cs.utas.edu.au)
Department of Computer Science, University of Tasmania
GPO Box 252C, Hobart, Tasmania 7001, Australia

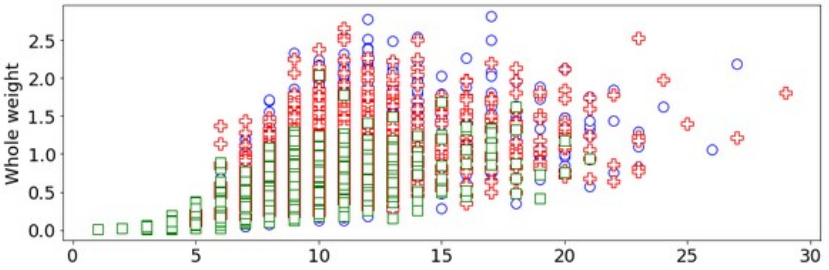
https://archive.ics.uci.edu/ml/datasets/Abalone

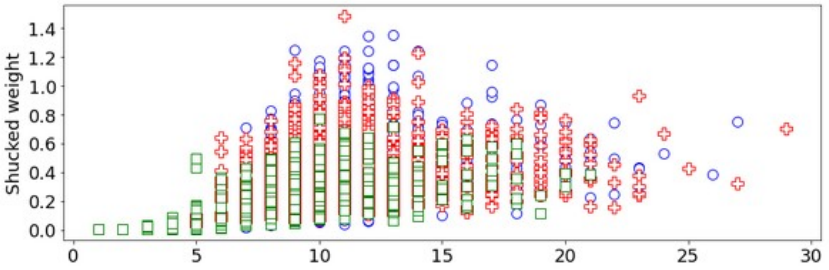

Sex
I
F
M
0
5
10
15
20
25
30

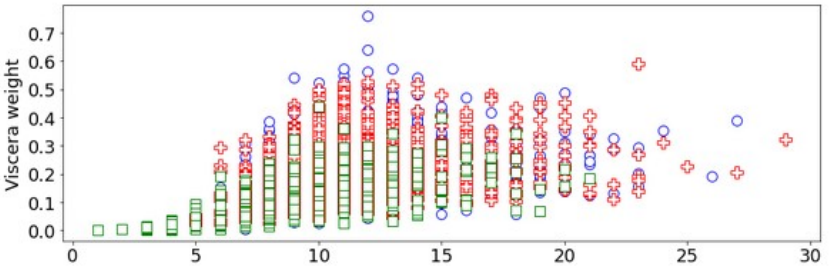

Length
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30

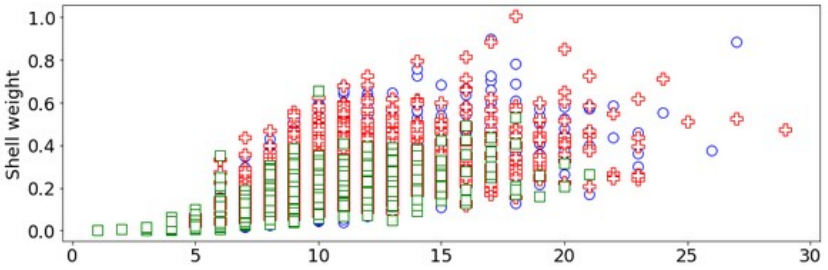

Diameter
0.6
0.5
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30


Height
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30


Whole weight
2.5
2.0
1.5
1.0
0.5
0.0
0
5
10
15
20
25
30


Shucked weight
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30


Viscera weight
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30

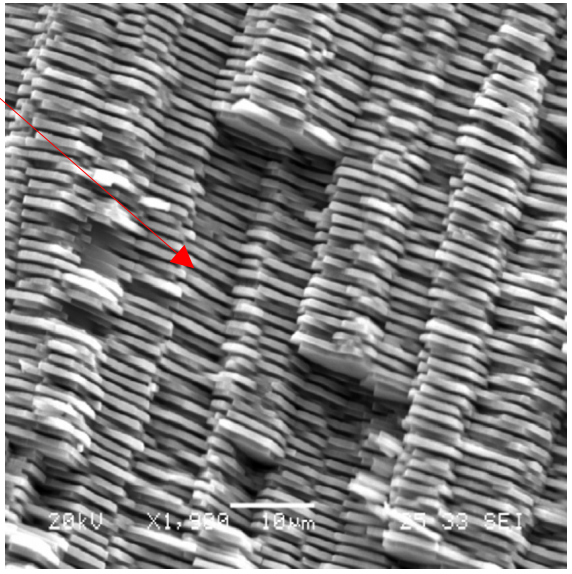

Shell weight
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30

Vista particular de los datos...

	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings	Years
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.1500	15	16.5
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.0700	7	8.5
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.2100	9	10.5
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.1550	10	11.5
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.0550	7	8.5
...	...	...	...	...	...	...	...	...	...	...
4172	F	0.565	0.450	0.165	0.8870	0.3700	0.2390	0.2490	11	12.5
4173	M	0.590	0.440	0.135	0.9660	0.4390	0.2145	0.2605	10	11.5
4174	M	0.600	0.475	0.205	1.1760	0.5255	0.2875	0.3080	9	10.5
4175	F	0.625	0.485	0.150	1.0945	0.5310	0.2610	0.2960	10	11.5
4176	M	0.710	0.555	0.195	1.9485	0.9455	0.3765	0.4950	12	13.5

4177 rows × 10 columns

The cross-section of the abalone shell showing a layered microstructure...

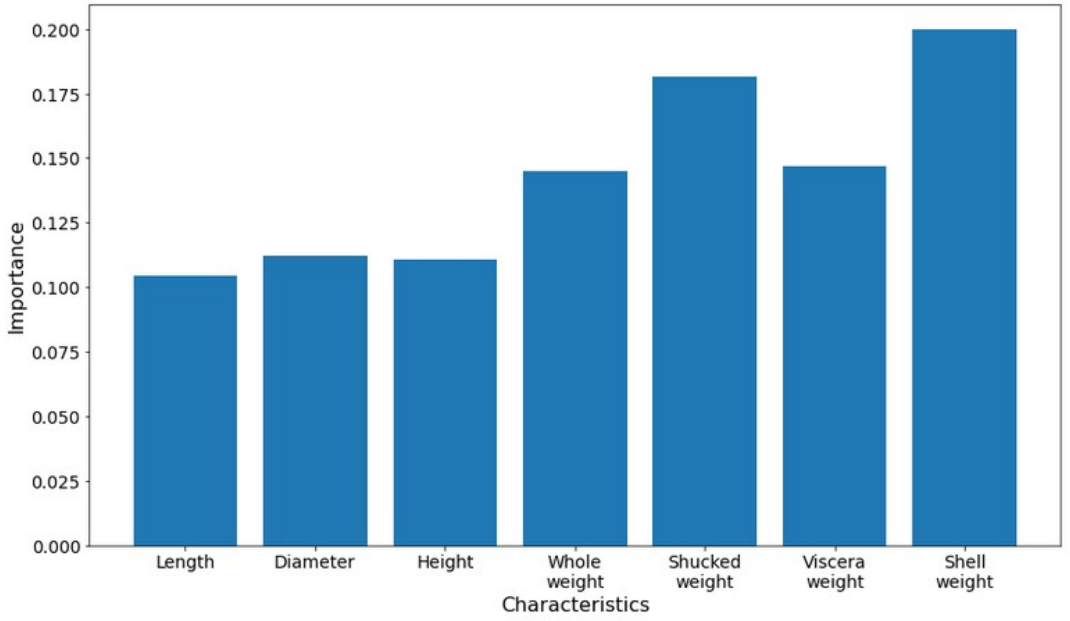


	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
Length	1.000000	0.986812	0.827554	0.925261	0.897914	0.903018	0.897706	0.556720
Diameter	0.986812	1.000000	0.833684	0.925452	0.893162	0.899724	0.905330	0.574660
Height	0.827554	0.833684	1.000000	0.819221	0.774972	0.798319	0.817338	0.557467
Whole weight	0.925261	0.925452	0.819221	1.000000	0.969405	0.966375	0.955355	0.540390
Shucked weight	0.897914	0.893162	0.774972	0.969405	1.000000	0.931961	0.882617	0.420884
Viscera weight	0.903018	0.899724	0.798319	0.966375	0.931961	1.000000	0.907656	0.503819
Shell weight	0.897706	0.905330	0.817338	0.955355	0.882617	0.907656	1.000000	0.627574
Rings	0.556720	0.574660	0.557467	0.540390	0.420884	0.503819	0.627574	1.000000

Matriz de correlación…

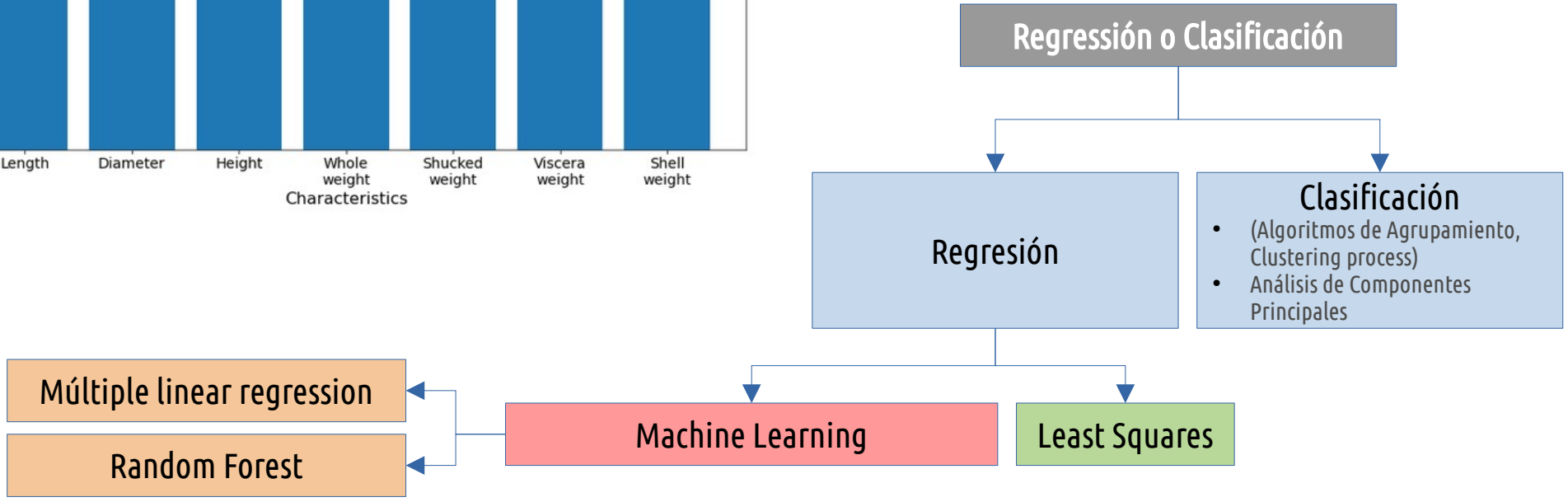
Algunas pares de variables están altamente correlacionadas, problemas de colinealidad...

```
Feature: Length, Score: 0.1
Feature: Diameter, Score: 0.11
Feature: Height, Score: 0.11
Feature: Whole weight, Score: 0.15
Feature: Shucked weight, Score: 0.18
Feature: Viscera weight, Score: 0.15
Feature: Shell weight, Score: 0.2
```



Importancia de las variables en un clasificador…

- Tenemos 7 variables independientes y 1 dependiente (que quiere ser predecida y/o replicada)
- **Fórmula:** Ring ~ Length + Diameter + Height + Whole weight + Shucked weight + Viscera weight + Shell weight
- Podemos ver el problema desde muchos puntos de vista:



Regresión → Least Squares

Minimizar una función llamada

Chi cuadrado = X²

$$E(\theta) = \chi^2 = \sum_{i=1}^{N} \left(\theta_j \cdot x_j^{(i)} + \theta_0 - y^{(i)}\right)^2$$

$$\begin{aligned}
\frac{\partial \chi^2}{\partial \theta_k} &= \frac{\partial}{\partial \theta_k} \sum_{i=1}^{N} \left(\theta_j \cdot x_j^{(i)} + \theta_0 - y^{(i)}\right)^2 \\
&= 2 \sum_{i=1}^{N} \left(\theta_j \cdot x_j^{(i)} + \theta_0 - y^{(i)}\right)^2 \frac{\partial}{\partial \theta_k} \left(\theta_j \cdot x_j^{(i)}\right) \\
&= 2 \sum_{i=1}^{N} \left(\theta_j \cdot x_j^{(i)} + \theta_0 - y^{(i)}\right)^2 x_k^{(i)} = 0
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \chi^2}{\partial \theta_0} &= \frac{\partial}{\partial \theta_0} \sum_{i=1}^{N} \left(\theta_j \cdot x_j^{(i)} + \theta_0 - y^{(i)}\right)^2 \\
&= 2 \sum_{i=1}^{N} \left(\theta_j \cdot x_j^{(i)} + \theta_0 - y^{(i)}\right)^2 \frac{\partial}{\partial \theta_0} (\theta_0) \\
&= 2 \sum_{i=1}^{N} \left(\theta_j \cdot x_j^{(i)} + \theta_0 - y^{(i)}\right)^2 = 0
\end{aligned}$$

Entonces el problema se reduce a calcular los coeficientes **θ** a partir de la solución del sistema de ecuaciones lineales

Variable	Coeficient
Length	-1.36813385
Diameter	13.2543506
Height	11.50289186
Whole weight	9.52350377
Shucked weight	-20.4315433
Viscera weight	-10.18253377
Shell weight	8.19106477
Intercept	2.964770977

Resultado final:

- **Random Forest** consigue un modelo que mejora el rendimiento y también la capacidad de predicción para el número de anillos en un 93% (según RMSLE) y en un 780% (según r²-score).

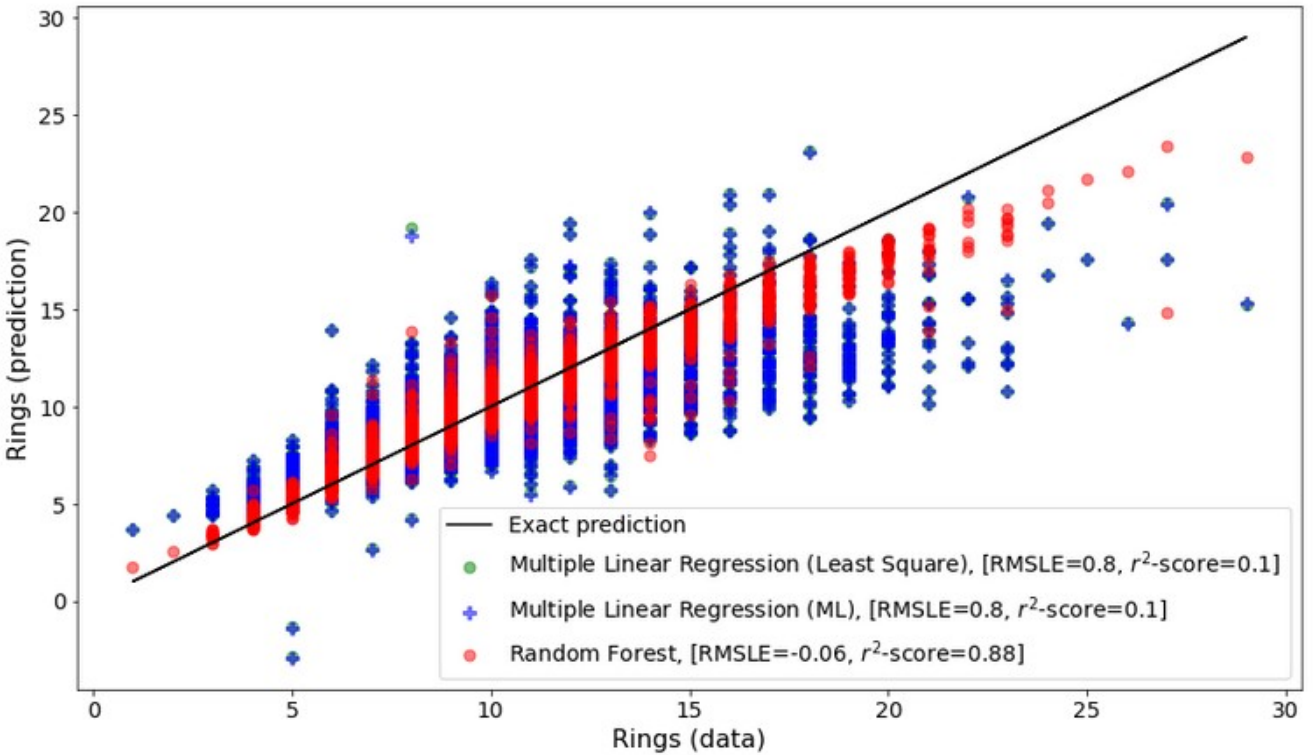
- Para evaluar el desempeño del modelo se usan varias métricas

- **Root Mean Squared Log Error (RMSLE)**

$$RMSLE = \sqrt{\frac{\sum_i \left(\log \frac{\hat{y}+1}{y+1}\right)^2}{n}}$$

- **R Squared (R2)**

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y}_i)^2}$$



https://github.com/sierraporta/Defensa_UTB