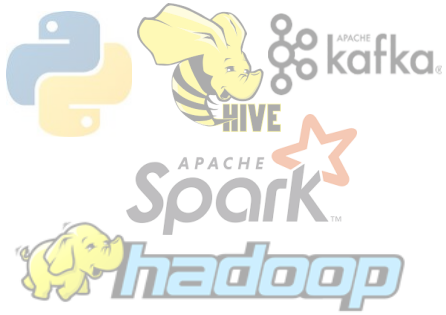

BIG DATA & ANALYTICS
Program Learning

Programa Integral en BigData & Analytics

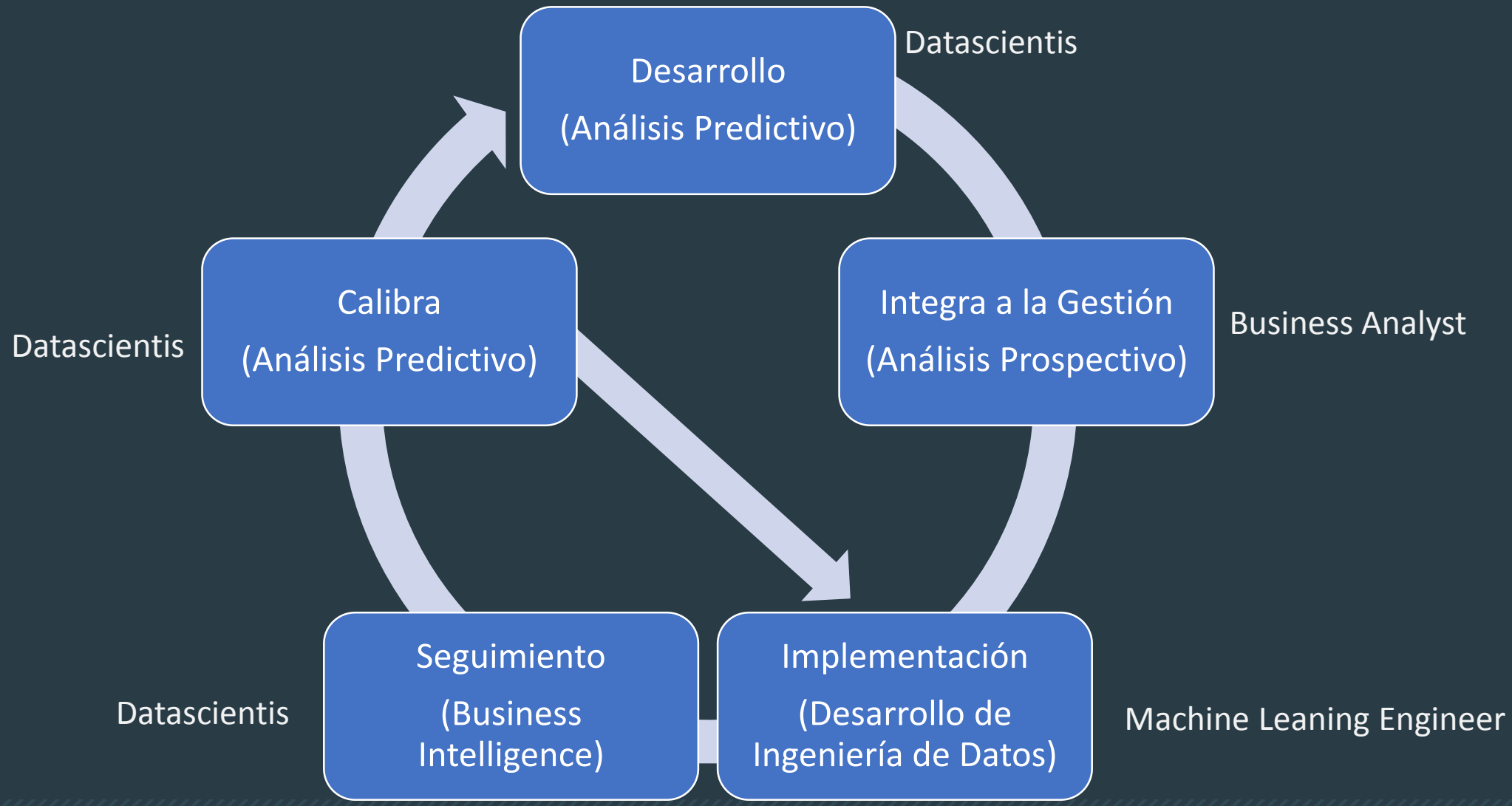
Business Intelligence (30hr)	Advanced Analytics (40hr)	Bigdata Tecnologies (42hr)	Agile Data (12hr)	Workshops (12hr)
DataWareHouse	Business Analytics	Big Data Fundamentals	SCRUM Data	Analytics with Unstructure Data
SQL Essentials	R Essentials	Python Essentials	DataOps	
Data Modeling	Modeling Process	Tecnologías Cloud y OnP	Data Transformations	
Data Engineering	Deploy and Monitoring	Storage technologies		
Data Visualization	Machine Learning Intro	Batch/RealTime Process		

Modeling Process

" … Generando valor a través de los Datos "

Proceso de Gestión de Modelos Predictivos (Ciclo de Vida)

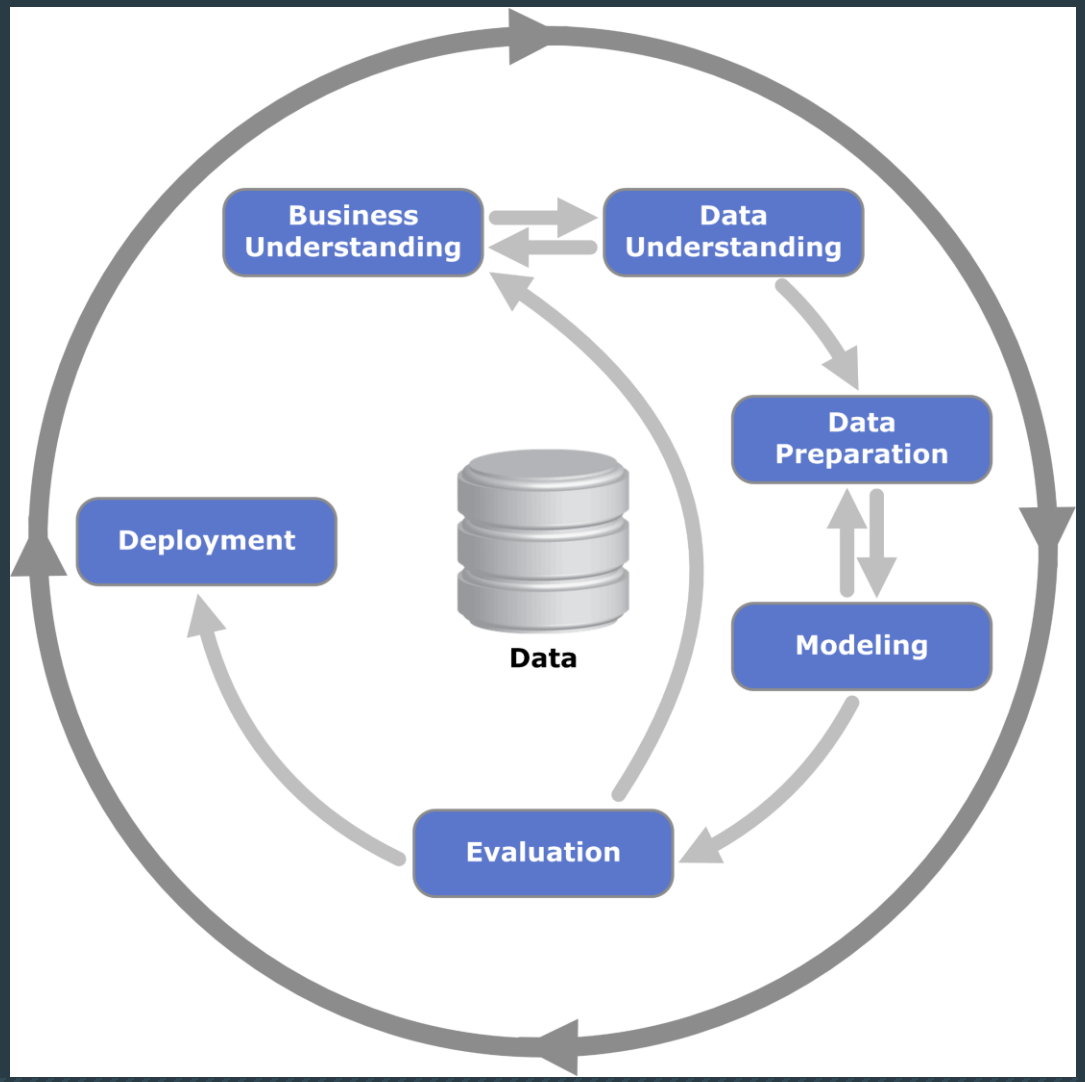


Fase de Desarrollo

Analítica Predictiva

Desarrollo (Proceso de Modelamiento)

- CRISP-DM (del inglés Cross Industry Standard Process for Data Mining)
- Se trata de un modelo estándar abierto del proceso que describe los enfoques comunes que utilizan los expertos en minería de datos.
- Es el modelo analítico más usado.



1. Business Understanding

- Se debe dedicar tiempo a explorar las expectativas de su organización con respecto al trabajo analítico.
- Intente implicar a la mayor cantidad de personas que sea posible en estas discusiones y documente los resultados.
- El paso final de la fase de CRISP-DM trata de cómo producir un plan de proyecto utilizando la información que se contiene en esta documentación.
- Aunque este estudio pueda parecer prescindible, no lo es. Conozca las razones comerciales para que sus esfuerzos aseguren que todos los usuarios están de acuerdo antes de asignar recursos.

1. Business Understanding : Caso de Uso

- La empresa de servicios financieros, pretende gestionar los eventos de fuga de clientes de manera que pueda oportunamente tomar acciones que beneficien al negocio, reteniendo a los clientes mas rentables y gestionando el inminente riego de fuga.
- Objetivo: Predicción de la fuga de clientes en una institución financiera, en una ventana que permita la gestión
- Criterios de Éxito:
  - Reducir la fuga en un 10%
  - Identificar a los clientes mas rentables
  - El estudio se completa dentro del plazo y presupuesto

2. Data Understanding

Cosiste en:

- La fase de comprensión de datos implica estudiar más de cerca los datos disponibles.
- Este paso es esencial para evitar problemas inesperados durante la siguiente fase (preparación de datos) que suele ser la fase más larga de un proyecto.
- La comprensión de datos implica acceder a los datos y explorarlos con la ayuda de tablas y gráficos.
- De esta forma podrá determinar la calidad de los datos y describir los resultados de estos pasos en la documentación del proyecto.

Actividades

- Recopilación de los datos iniciales
- Descripción de los datos
- Exploración de los Datos
- Verificación de la calidad de los datos

Entregales

- Informe del diagnostico de Datos
- Planteamiento de segmentaciones
- Separación de la data:
  - Muestra para el desarrollo
  - Fuera de Muestra (OOS)

2. Data Understanding: Recopilación

- Para el caso de uso, “Fuga de clientes” se cuenta con la siguiente data:

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
2	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
4	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
5	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0
6	6	15574012	Chu	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
7	7	15592531	Bartlett	822	France	Male	50	7	0.00	2	1	1	10062.80	0
8	8	15656148	Obinna	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1

Showing 1 to 9 of 10,000 entries

2. Data Understanding: Descripción

- Existen muchas formas de describir datos, pero por lo general esta actividad se centra en la cantidad y calidad de los datos; la cantidad de datos disponible y el estado de los datos.

```
> summary(df)
   RowNumber        CustomerId        Surname      CreditScore      Geography      Gender          Age        
 Min.   :    1   Min.   :15565701   Smith   :  32   Min.   :350.0   France :5014   Female:4543   Min.   :18.00  
 1st Qu.: 2501   1st Qu.:15628528   Martin  :  29   1st Qu.:584.0   Germany:2509   Male  :5457   1st Qu.:32.00  
 Median : 5000   Median :15690738   Scott   :  29   Median :652.0   Spain  :2477                 Median :37.00  
 Mean   : 5000   Mean   :15690941   Walker  :  28   Mean   :650.5                                Mean   :38.92  
 3rd Qu.: 7500   3rd Qu.:15753234   Brown   :  26   3rd Qu.:718.0                                3rd Qu.:44.00  
 Max.   :10000   Max.   :15815690   Genovese:  25   Max.   :850.0                                Max.   :92.00  
                                    (Other) :9831                                                               
     Tenure          Balance       NumOfProducts      HasCrCard        IsActiveMember   EstimatedSalary        Exited      
 Min.   : 0.000   Min.   :     0   Min.   :1.00   Min.   :0.0000   Min.   :0.0000   Min.   :    11.58   Min.   :0.0000  
 1st Qu.: 3.000   1st Qu.:     0   1st Qu.:1.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 50983.75   1st Qu.:0.0000  
 Median : 5.000   Median : 97199   Median :1.00   Median :1.0000   Median :1.0000   Median :100252.26   Median :0.0000  
 Mean   : 5.013   Mean   : 76486   Mean   :1.53   Mean   :0.7055   Mean   :0.5151   Mean   :100099.21   Mean   :0.2037  
 3rd Qu.: 7.000   3rd Qu.:127644   3rd Qu.:2.00   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:149400.92   3rd Qu.:0.0000  
 Max.   :10.000   Max.   :250898   Max.   :4.00   Max.   :1.0000   Max.   :1.0000   Max.   :199992.48   Max.   :1.0000  
                                                                                     NA's   :6                           
```

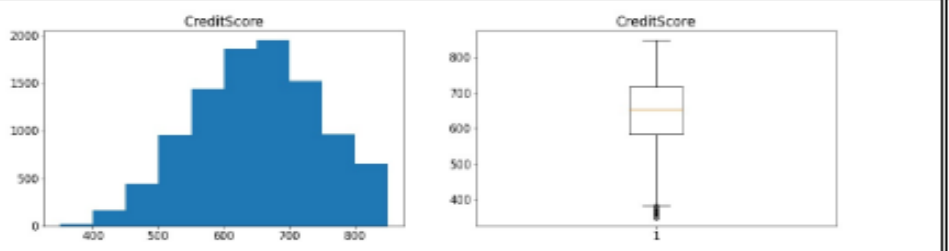
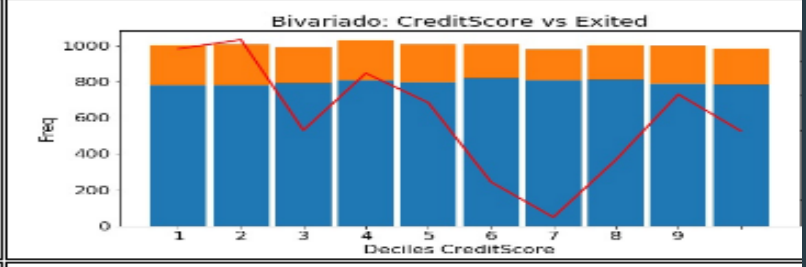
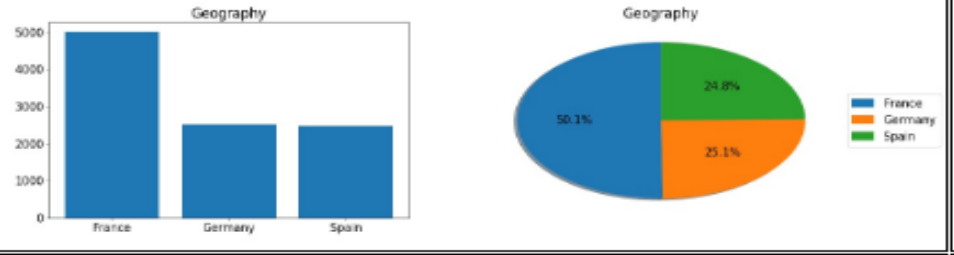
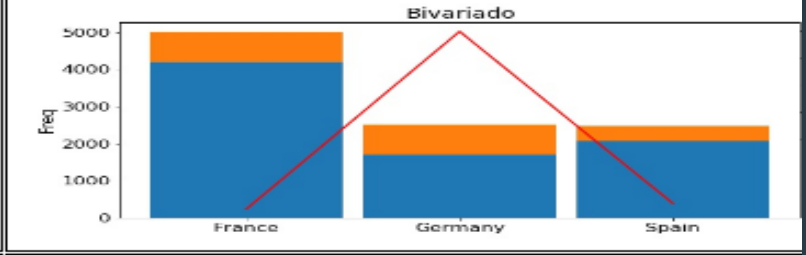
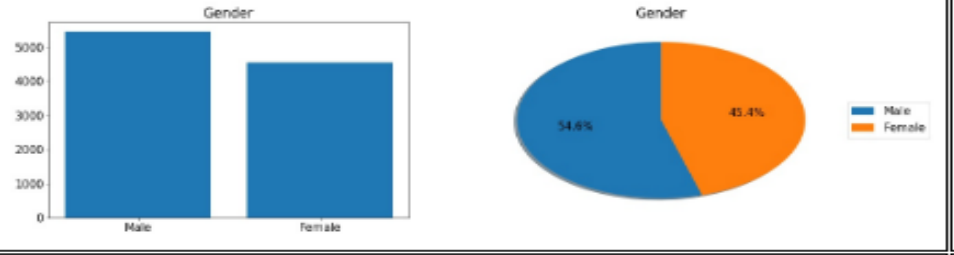
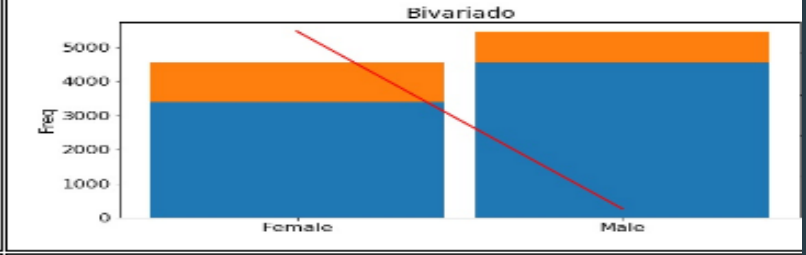
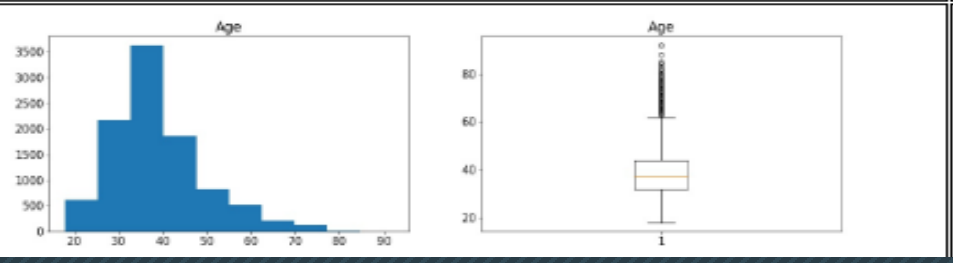
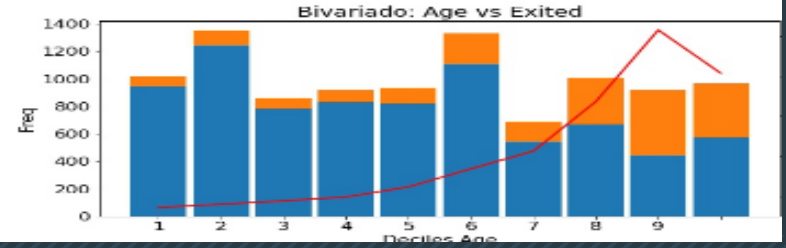
2. Data Understanding: Exploración

- Estos análisis ayudan a descubrir el sentido de los datos de cara al negocio y al problema a resolver.
- También pueden ayudarle a formular hipótesis y dar forma a las tareas de transformación de datos.
- Utilice esta fase para explorar los datos con las tablas, gráficos y otras herramientas de visualización

2. Data Understanding: Exploración

Tipos de Datos	Técnicas (Distribución del los Datos)	
	Visuales	Numéricas
Cualitativos	Bar plot Pie plot	Frecuencias Absolutas y Relativas, simples o acumuladas
Cuantitativos	Hist Box plot	Medidas de Tendencia Central Medidas de Dispersión Medidas de Posición

2. Data Understanding: Exploración

Nro	ESTADISTICOS	UNIVARIADO	BIVARIADO
1	Nmiss 0.0 Q1 584.0 Q3 718.0 cv 0.14857650996563196 k -0.42611279176518124 max 850 mean 650.5288 median 652.0 min 350 mode 850 sd 96.65329873613061 var CreditScore		
2	Nmiss . Q1 . Q3 . cv . k . max . mean . median . min . mode . sd . var Geography		
3	Nmiss . Q1 . Q3 . cv . k . max . mean . median . min . mode . sd . var Gender		
4	Nmiss 0.0 Q1 32.0 Q3 44.0 cv 0.2694584128099058 k 1.3940495456392599 max 92 mean 38.9218 median 37.0 min 18 mode 37 sd 10.487806451704591 var Age		

2. Data Understanding: Segmentación

Segmentar una población consiste dividir en diferentes grupos, de manera que cada grupo tenga características comunes

- Se busca homogeneidad intra-grupos y heterogeneidad inter-grupos
- Para esto se debe identificara variables segmentadoras (Drivers de segmentación)

Tras la segmentación la muestra total quedará separ4a en sub poblaciones. En lugar de realizar un modelo para la muestra total, se hará uno para cada sub población.

2. Data Understanding: Segmentación

Al segmentar una población, esta queda separada en varias subpoblaciones. El objetivo es modelizar cada subpoblación de manera independiente

- Generalmente, realizar una análisis independiente para cada subpoblación, permite obtener unos resultados mas precisos, que desarrollando el mismo análisis para la muestra conjunta.
- Eso suele ser así dado que comúnmente una población esta formada por diferentes subpoblaciones y un mismo modelo no funcionaría igual para todas ellas, debido a que cada subpoblación tiene unas características diferentes (realizar modelos por separado pude aumentar el poder predictivo de los mismos).

Poderes predictivos en modelos independientes

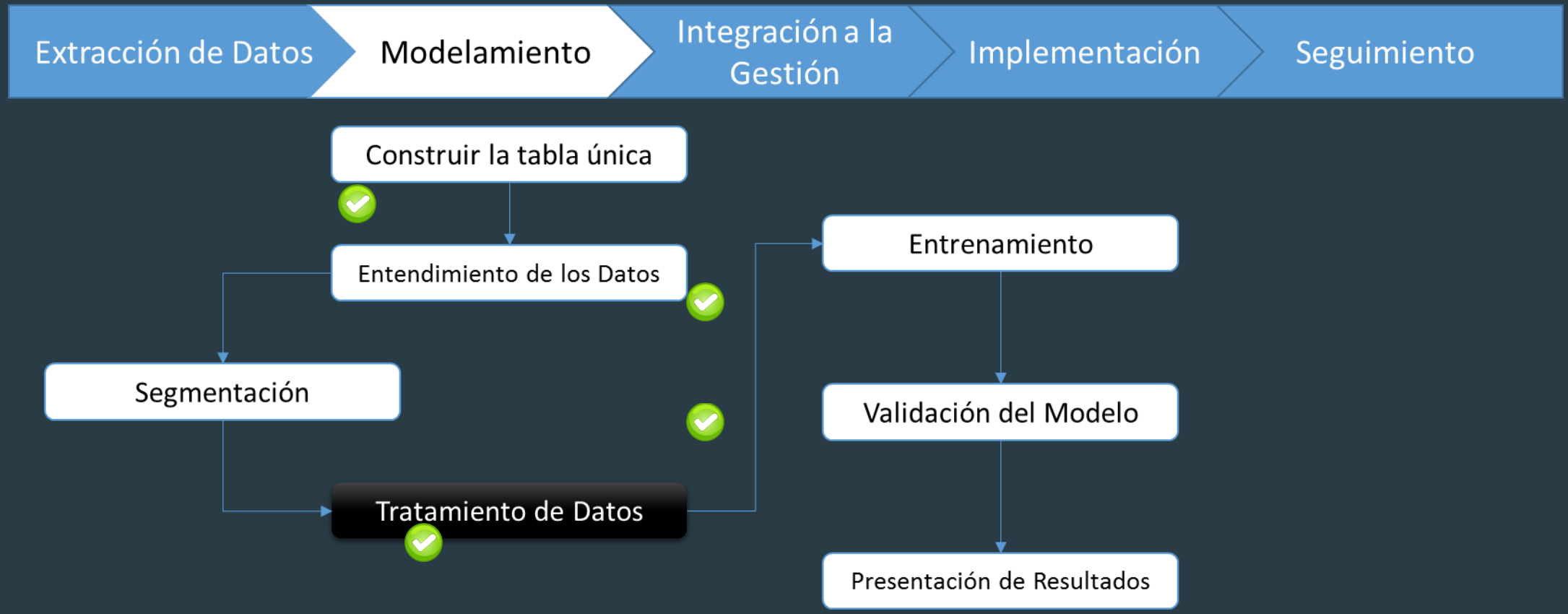
Eje	ROC
Subpoblación 1	84%
Subpoblación 2	83%
Subpoblación 3	85%

Poderes predictivos en modelos único y aplicado a las subpoblaciones

Eje	ROC	Eje	ROC
Modelo Único	81%	Subpoblación 1	80%
		Subpoblación 2	79%
		Subpoblación 3	74%

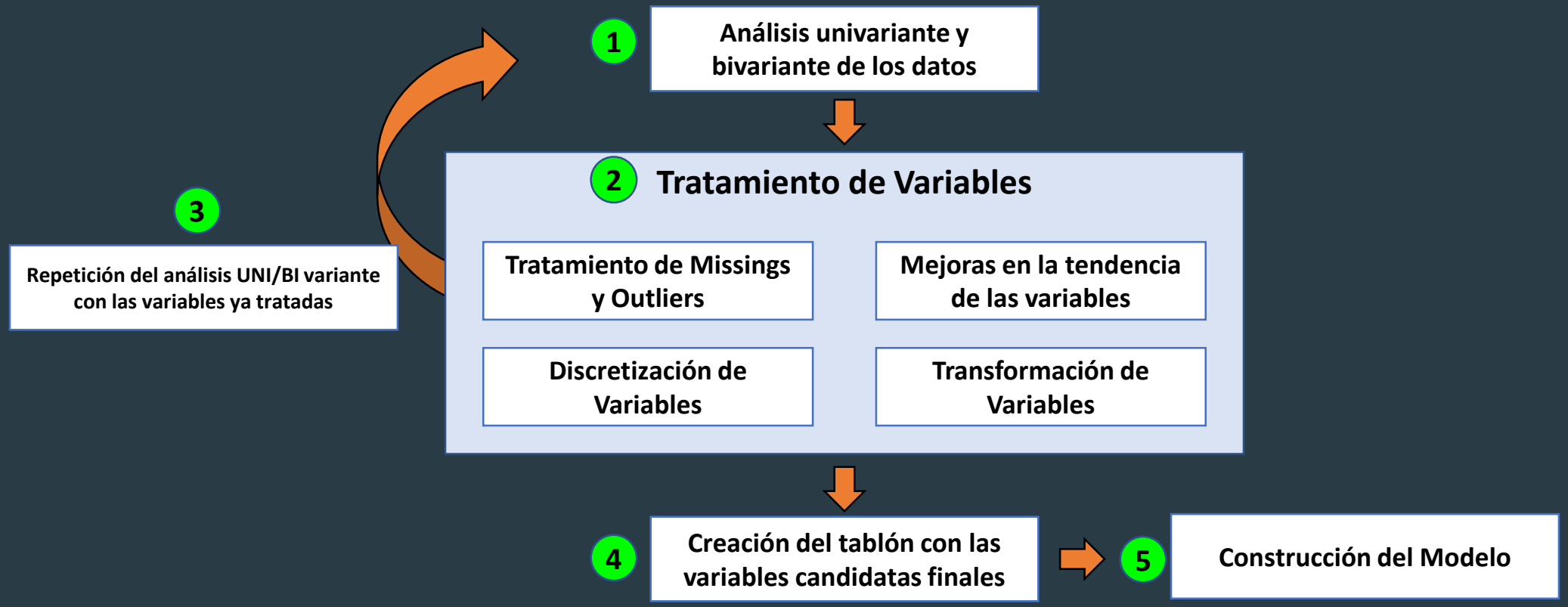
Realizar un modelo único con subpoblaciones muy heterogéneas, puede llevara a no predecir bien algunos ejes

3. Data Preparations



3. Data Preparations

A partir del análisis univariante y bivariante de la extracción se lleva a cabo un tratamiento de variables con el objetivo de perfilar los datos que finalmente serán utilizados en la construcción del modelo.



3. Data Preparations: Tratamiento de Datos

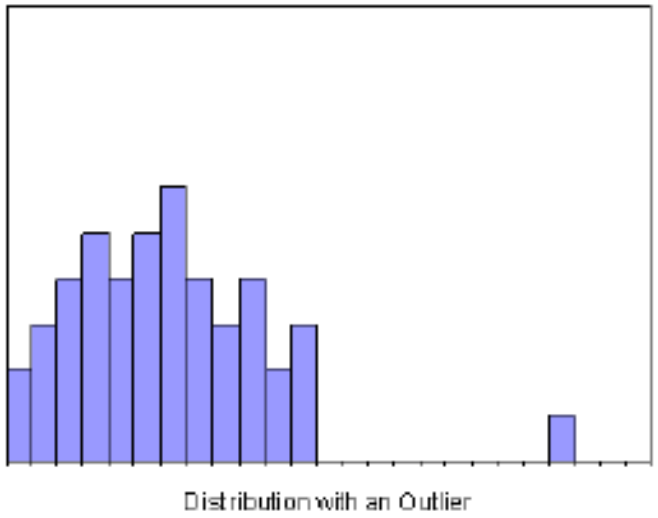
El tratamiento de variables tiene como objetivo la mejora de la calidad de los datos que serán utilizados en la construcción de modelos, utilizando métodos de tratamiento de missings y outliers, de mejora de la tendencia de variables, discretización y transformación de variables.

- **Análisis y tratamiento de missings** (valores no informados) **y outliiers** (atípicos, valores fuera de dominio). Algunas técnicas son:
  - Eliminación de registros
  - Sustitución por otro valor
  - Inclusión de dummies
  - Discretización de la variables
- **Creación de nuevas variables a partir de otras existentes en la tabla única**. Algunos ejemplos:
  - Scalamiento
  - Transformar fechas en antigüedad
  - Crear ratios, promedios, conteos, máximos, minimos, etc.

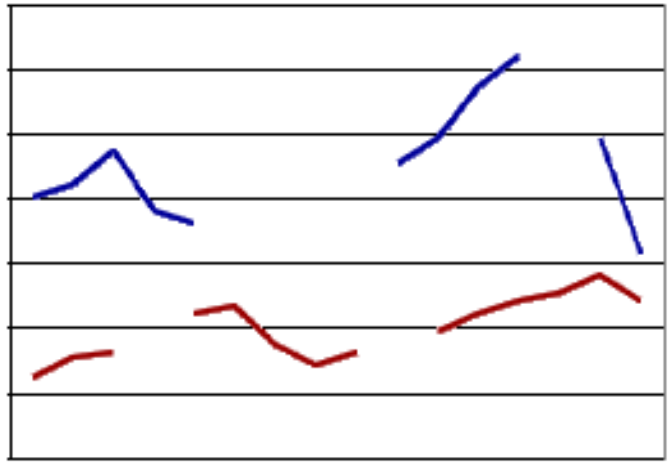
- **Introducir mejoras en la tendencia de las variables**.
  - Tratamiento del Target
  - Suavización de variables
  - Discretización
- **Creación de nuevas variables a partir de variables continuas, mediante la creación de grupos homogeneos**.
  - Creación de grupos de forma experta a partir de análisis de frecuencias o histogramas
  - Métodos de clustering
  - Discretización automática buscando heterogeneidad del target.

3.1 Tratamiento de Datos

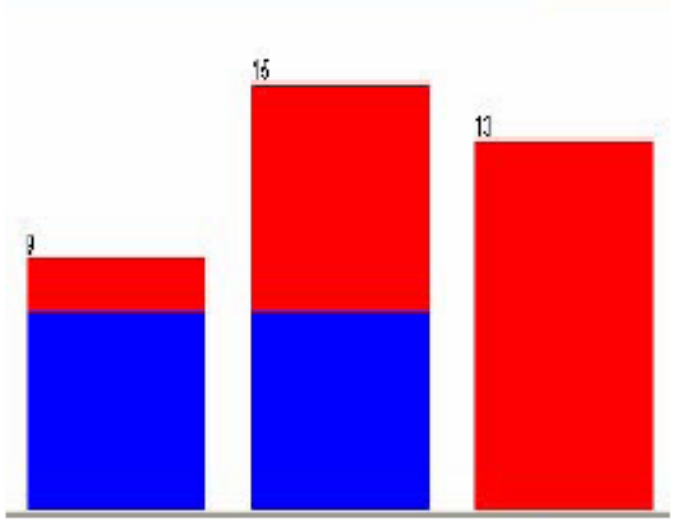
- Tratamiento de Missings y Outliers
- Este tipo de tratamiento mejora la calidad, distribución y tendencia de las variables. Se determinan a partir de los resultados de los análisis Univariantes y Bivariantes



Valores extremos



Valores null



inconsistentes

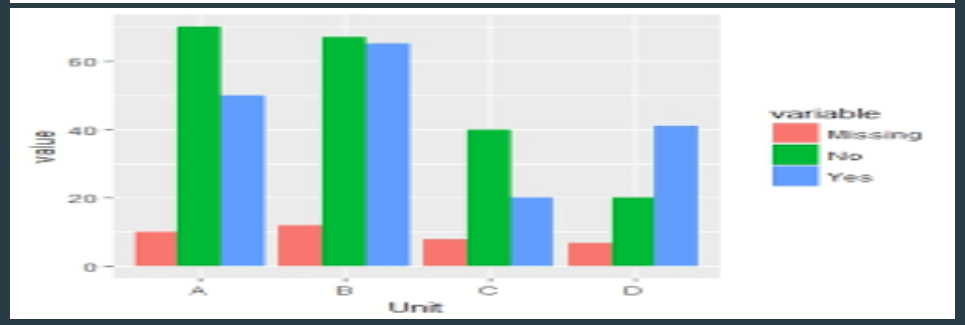
3.1 Tratamiento de Datos: Missings

- Identificación:
  - Antes del tratamiento de missings es preciso realizar un análisis experto de los motivos de la desinformación, esto permite determinar cual es la metodología de tratamiento mas idónea

Datos Discretos:

```
Frequencies
Type
Data frame: data
Type: Factor (unordered)
```

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
E	19	31.67	31.67	29.69	29.69
B	15	25.00	56.67	23.44	53.12
C	11	18.33	75.00	17.19	70.31
D	8	13.33	88.33	12.50	82.81
A	7	11.67	100.00	10.94	93.75
<NA>	4			6.25	100.00
Total	64	100.00	100.00	100.00	100.00



Datos Continuos:

```
     Ozone           Solar.R         Wind           Temp
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :57.00
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:73.00
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00
 Mean   : 42.13   Mean   :185.9   Mean   : 9.806   Mean   :78.28
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00
 NA's   :37       NA's   :7       NA's   :7        NA's   :5
```

3.1 Tratamiento de Datos: Missings

- Tratamiento:
  - Si el porcentaje de missings de una variable es relevante y se quiere mantener dicha variable en el modelo, es recomendable realizar un tratamiento para no perder un porcentaje elevado de la población. Es preciso establecer un umbral de missings aceptable

- Reemplazar (Recomendable)
  - Por una constante global para reconocimiento del algoritmo
  - Por la media del resto de observaciones (de la misma clase)
  - Por el valor mas probable obtenido por una técnica de inferencia (Bayes, Arboles, etc.)

- Pasar por alto (No recomendable)
  - Ignorar, Existen algoritmos robustos
  - Eliminar la variable
  - Filtrar las filas

A la hora de realizar el tratamiento es importante:

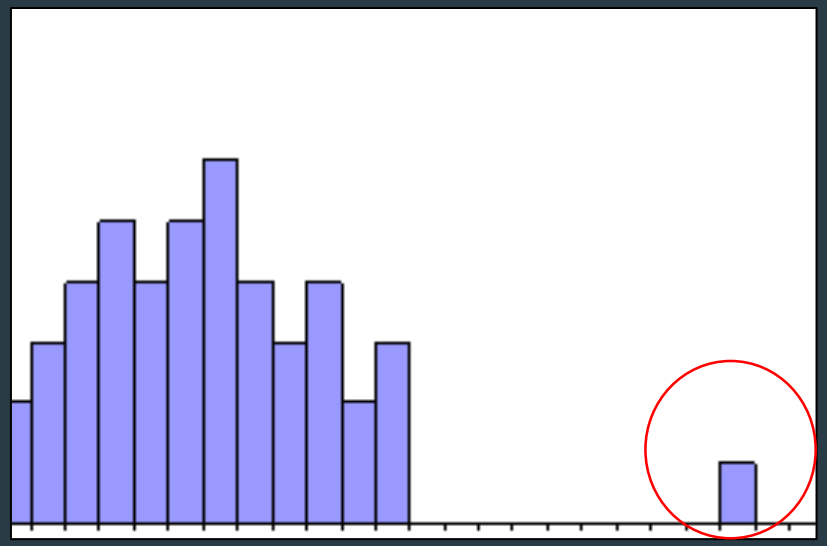
1. No modificar la distribución de los datos y
2. No cambiar la relación de la variable tratada con el resto de variables

3.1 Tratamiento de Datos: Outliers

Identificación:

Método Visual

- Diagrama de Frecuencias
- Diagrama de Cajas (Boxplot)



Método Analítico

- Identificar puntos que se encuentra fuera de la $\mu \pm 3\sigma$
- Las que están fuera del Rango intercuartílico :

$[Q1 - 1.5 * IRQ , Q3 + 1.5 * IRQ]$

3.1 Tratamiento de Datos: Outliers

Tratamiento:

Recomendable

- Discretizar. Transformar un valor continuo en discreto (Muy alto, …, muy Bajo)
- Imputar Percentiles extremos (Pe. P1 y P99)

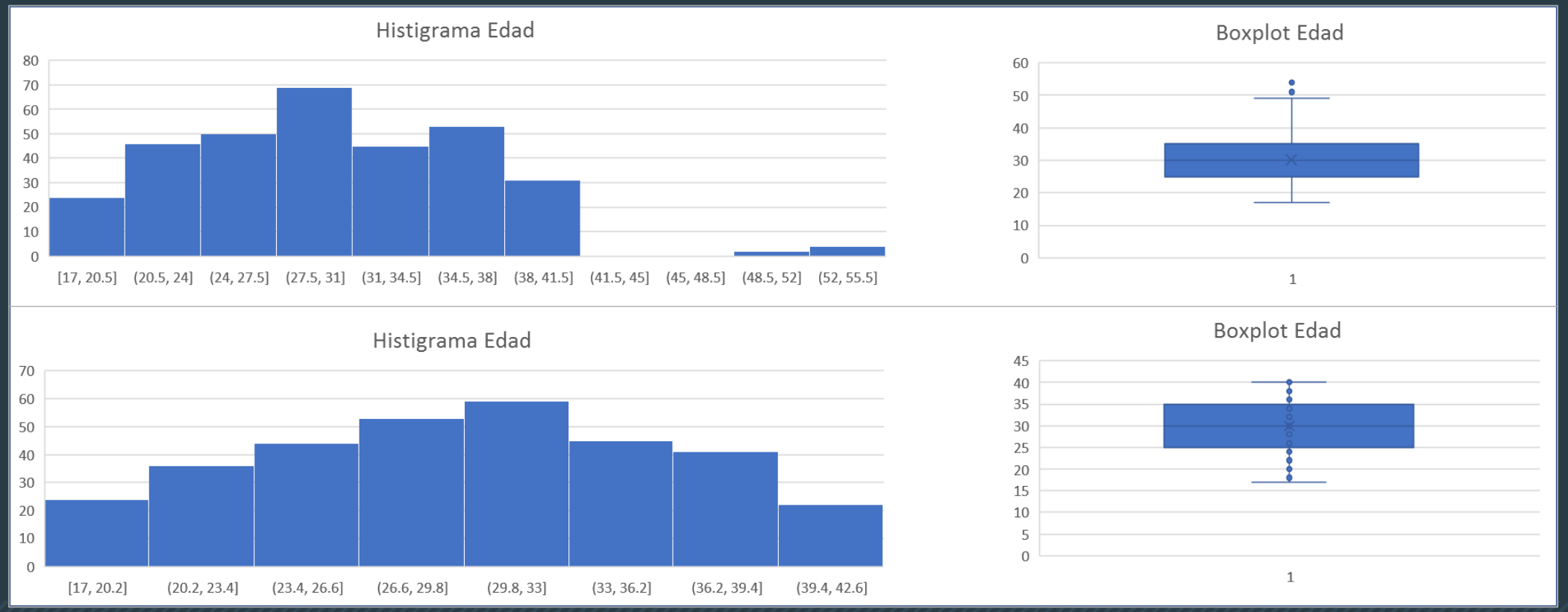
No recomendable

- Ignorar, Existen algoritmos robustos
- Eliminar la variable
- Filtrar las filas

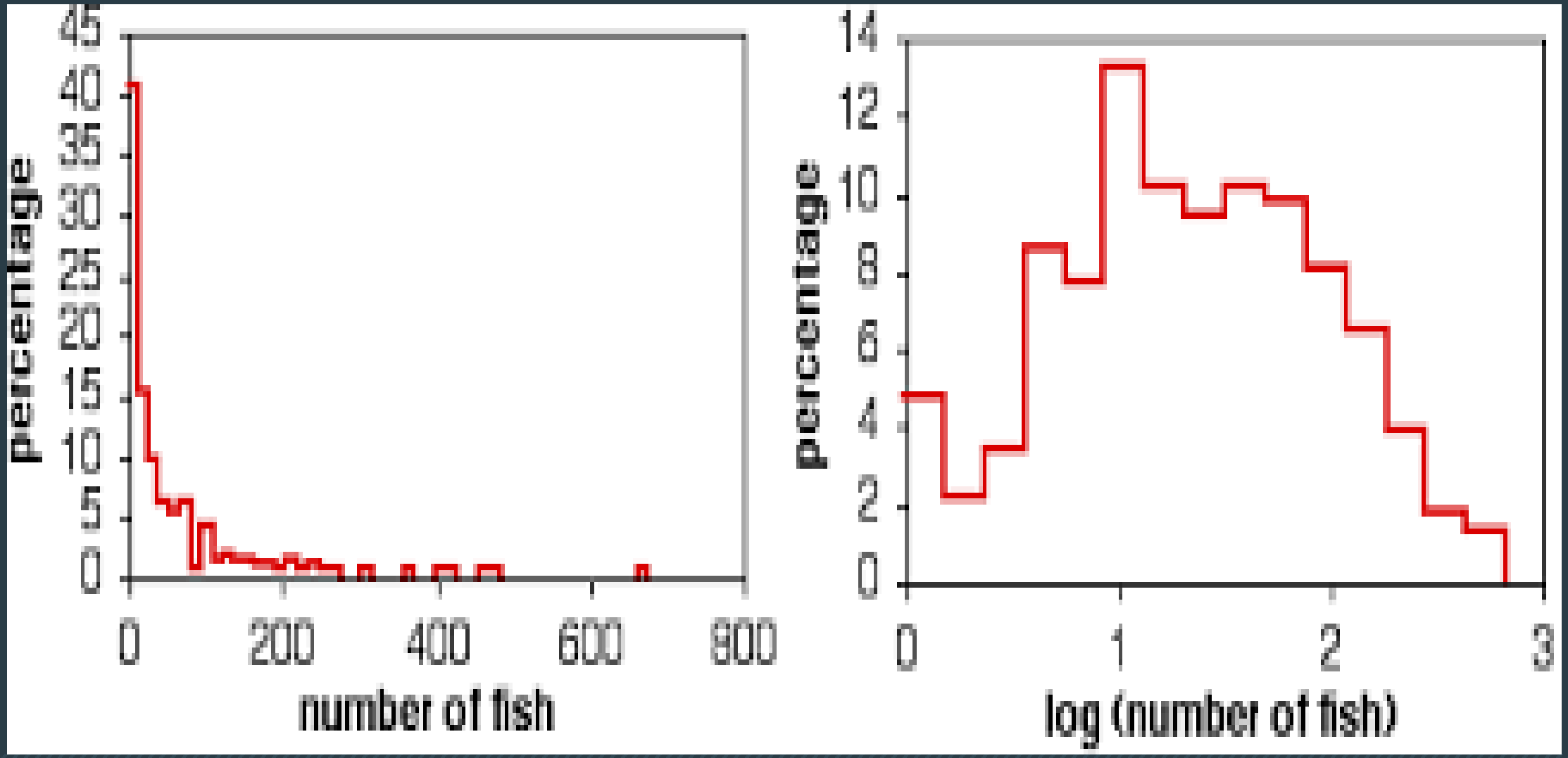
3.1 Tratamiento de Datos: Outliers

Tratamiento:

Un atípico es una observación alejada de la mediana y cuya eliminación de la muestra podría suponer un cambio significativo en la estimación del modelo. Su tratamiento no debería distorsionar la distribución inicial de la población.



3. Data Preparations: Transformación de Datos

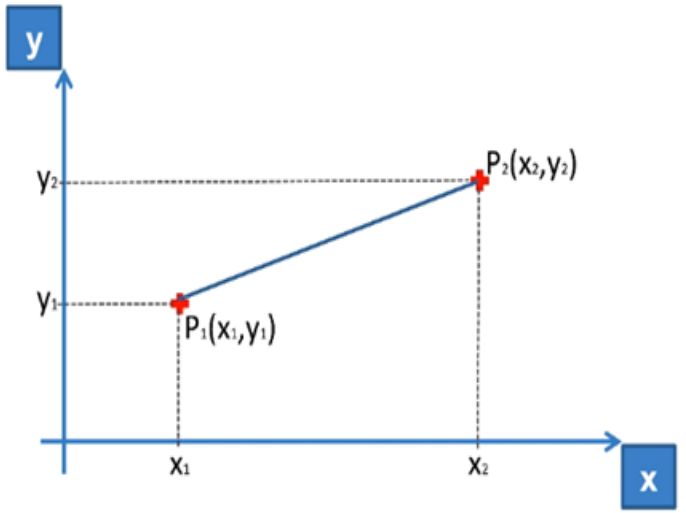


3.2 Transformación de Datos

Feature Scaling

- La mayoría de algoritmos de aprendizaje automático, se basa en lo que se llama la distancia euclidiana. Tener valores en diferente escala causará problemas en este calculo

	Country	Age	Salary	Purchased
2	France	44	72000	No
3	Spain	27	48000	Yes
4	Germany	30	54000	No
5	Spain	38	61000	No
6	Germany	40	63777.77778	Yes
7	France	35	58000	Yes
8	Spain	38.77777778	52000	No
9	France	48	79000	Yes
10	Germany	50	83000	No
11	France	37	67000	Yes



Euclidean Distance between P_1 and $P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Se recomienda tener las variables en el mismo rango, en la misma escala, de modo que ninguna variable esté dominada por otra.

3.2 Transformación de Datos

Feature Scaling

- Existen varias formas de escalar los datos, las mas comunes son las siguientes:

Standardisation	Normalisation
$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$	$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$

- Transformando las variables de valores grandes y muy diferentes a valores pequeños y similares.
- Incluso si los algoritmos no están basados en la distancia euclidiana, es conveniente hacer el escalamiento para que el algoritmo converja mucho mas rápido (Arboles de decisión).
- El escalamiento de variables categóricas no es recomendable dado que perdemos interpretación del modelo resultante.

3.2 Transformación de Datos

Feature Scaling

	Country	Age	Salary	Purchased
1	1	44.00000	72000.00	0
2	2	27.00000	48000.00	1
3	3	30.00000	54000.00	0
4	2	38.00000	61000.00	0
5	3	40.00000	63777.78	1

```
# Feature Scaling
dataset[,2:3] = scale(dataset[,2:3])
```

	Country	Age	Salary	Purchased
1	1	0.7199314	0.7110128	0
2	2	-1.6236751	-1.3643758	1
3	3	-1.2100975	-0.8455287	0
4	2	-0.1072238	-0.2402070	0
5	3	0.1684946	0.0000000	1

Considerar lo siguiente, respeto al escalamiento de la variable target:

- Modelos de regresión: recomendable
- Modelos de clasificación: no escalar la variable target

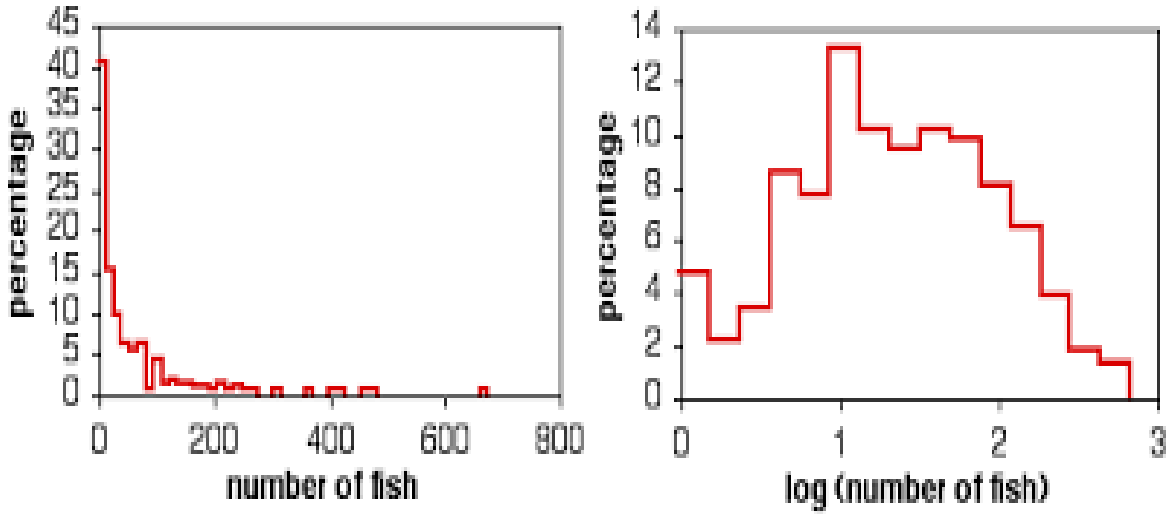
3.2 Transformación de Datos

Feature Transformations

- Un forma muy común de reducir la escala de una variables es utilizando el logaritmo. A este procedimiento también se le conoce como suavización de datos. Se aplica cuando se tiene valores numéricos muy grandes y de alta variabilidad

Logaritmo natural, en base 2 o en base 10.

$$v' = \log_{10}(v)$$



3.2 Transformación de Datos

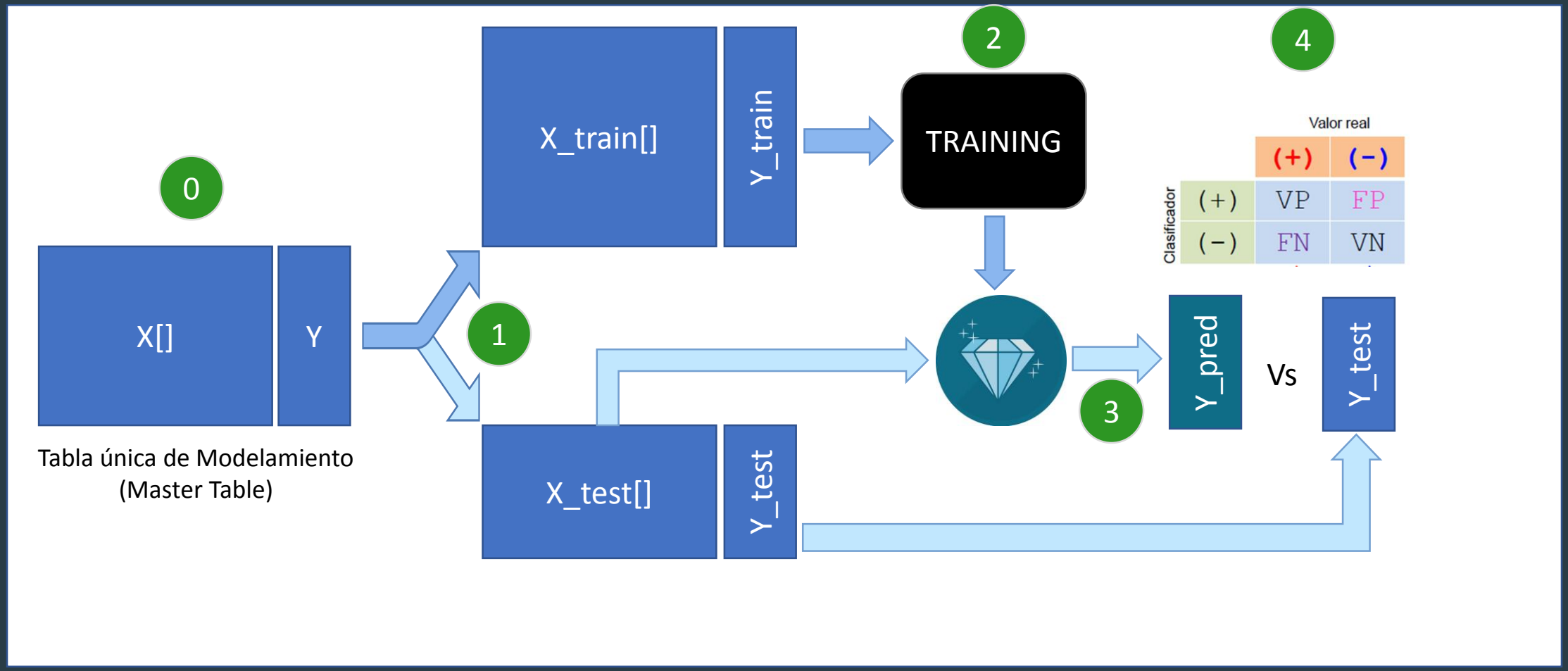
Encoding categorical Data

- Algunos algoritmos requieren que todas las variables sean numéricas, por lo que debemos recodificar las variables categóricas.

```
# Encoding categorical data
dataset$Country = factor(dataset$Country,
                         levels = c('France', 'Spain', 'Germany'),
                         labels = c(1, 2, 3))
dataset$Purchased = factor(dataset$Purchased,
                           levels = c('No', 'Yes'),
                           labels = c(0, 1))
```

4. Modeling

Workflow de Modelamiento



4. Modeling: Split Train Test

Dividir la data en dos muestras, es una técnica muy utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de las dos muestras (cross-validation)

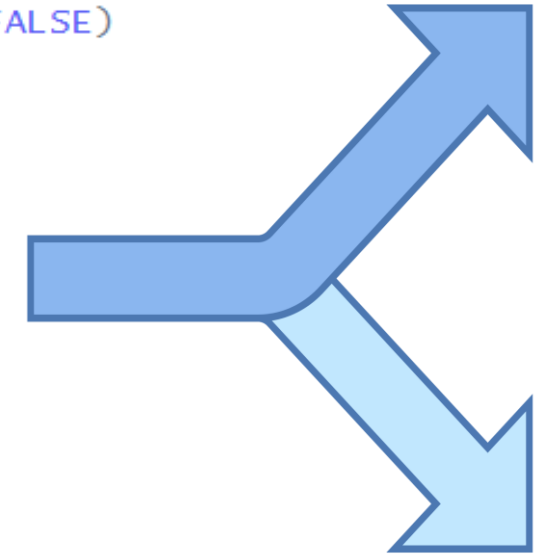
- La muestra de train, será con la cual se realiza el entrenamiento y como consecuencia se obtendrá una predicción (formula o regla lógica)
- La muestra de test, permitirá evaluar la predicción. Aplicando la regla obtenida, a este nuevo set de datos (target real vs target estimado)
- Solo aplica para aprendizaje supervisado.
- El muestreo tiene que estar balanceado por la variable target.
- La división de las observaciones recomendada es:
  - training 60-80% del total
  - test 40-20% del total

4. Modeling: Split Train Test

El porcentaje de división dependerá del algoritmo seleccionado

```
# install.packages('caTools')
library(caTools)
set.seed(123)
split = sample.split(dataset$Purchased, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)
```

	Country	Age	Salary	Purchased
1	1	44.00000	72000.00	0
2	2	27.00000	48000.00	1
3	3	30.00000	54000.00	0
4	2	38.00000	61000.00	0
5	3	40.00000	63777.78	1
6	1	35.00000	58000.00	1
7	2	38.77778	52000.00	0
8	1	48.00000	79000.00	1
9	3	50.00000	83000.00	0
10	1	37.00000	67000.00	1



	Country	Age	Salary	Purchased
1	1	44.00000	72000.00	0
2	2	27.00000	48000.00	1
3	3	30.00000	54000.00	0
4	2	38.00000	61000.00	0
5	3	40.00000	63777.78	1
7	2	38.77778	52000.00	0
8	1	48.00000	79000.00	1
10	1	37.00000	67000.00	1

	Country	Age	Salary	Purchased
6	1	35	58000	1
9	3	50	83000	0

4. Modeling: Algoritmos de Aprendizaje Predictivo

Clasificación

- Predice una categoría (nro finito de valores) o la probabilidad de pertenencia a la categoria.
- Ejemplos:
  - ¿comprará el cliente este producto? [sí, no]
  - ¿tipo de tumor? [maligno, benigno]
  - ¿subirá el índice bursátil? IBEX mañana [sí, no]
  - ¿nos devolverá este cliente un crédito? [sí, no]
  - ¿qué deporte estás haciendo? tal y como lo detectan los relojes inteligentes [caminar, correr, bicicleta, nadar]

Regresión

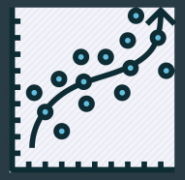
- Predice un valor numérico (valor numérico dentro de un conjunto infinito de posibilidades)
- Ejemplos:
  - Predecir por cuánto se va a vender una propiedad inmobiliaria
  - Predecir cuánto tiempo va a permanecer un empleado en una empresa
  - Estimar cuánto tiempo va a tardar un vehículo en llegar a su destino
  - Estimar cuántos productos se van a vender

4. Modeling: Algoritmos de Aprendizaje Predictivo

Clasificación

- Hay varias Algoritmos que podemos usar en problemas de clasificación. Por ejemplo:
  - **Logistic regression (Sólo funciona para clasificar)**
  - Support vector machines
  - **Decision trees**
  - Random forests
  - Redes neuronales
  - …

Regresión



- Aunque hay algunas técnicas que son específicas de clasificación y otras de regresión, la mayoría de las técnicas funcionan con ambos:
  - **Regresión lineal y no lineal**
  - Support vector machines
  - **Decision trees**
  - Random forests
  - **Redes neuronales**
  - …

4. Modeling: Algoritmos de Aprendizaje Predictivo