UNIVERSIDAD
NACIONAL
DE COLOMBIA

APRENDIZAJE DE MÁQUINAS

JOHN W. BRANCH

Profesor Titular
Departamento de Ciencias de la Computación y de la Decisión
Director del Grupo de I+D en Inteligencia Artificial – GIDIA
jwbranch@unal.edu.co

https://github.com/srobles05/3008422-AprendizajeDeMaquinas

https://www.coursera.org/programs/universidad-nacional-de-colombia-odgoe

Gidia Grupo de I+D en Inteligencia Artificial

UNIVERSIDAD NACIONAL DE COLOMBIA

METODOLOGÍA ENSEÑANZA – APRENDIZAJE

Sesiones Remotas vía Google.Meet

Sincrónicas y Asincrónicas

El **aprendizaje sincrónico** involucra estudios online a través de una plataforma. Este tipo de aprendizaje sólo ocurre en línea. Al estar en línea, el estudiante se mantiene en contacto con el docente y con sus compañeros. Se llama aprendizaje sincrónico porque la plataforma permite que los estudiantes pregunten al docente o compañeros de manera instantánea a través de herramientas como el chat o el video chat.

El **aprendizaje asincrónico** puede ser llevado a cabo online u offline. El aprendizaje asincrónico implica un trabajo de curso proporcionado a través de la plataforma o el correo electrónico para que el estudiante desarrolle, de acuerdo a las orientaciones del docente, de forma independiente. Un beneficio que tiene el aprendizaje asincrónico es que el estudiante puede ir a su propio ritmo.

Descripción del Curso

El curso introduce los conceptos fundamentales y los métodos más utilizados en el campo del aprendizaje de máquinas enfocados desde las perspectivas de la naturaleza del problema que se requiere resolver, esto es, aprendizaje supervisado orientado a los problemas de clasificación y regresión para aplicaciones de predicción o pronóstico. Aprendizaje no supervisado orientado a tareas de agrupar o etiquetar un conjunto de datos, También se incluyen la aproximación general de técnicas modernas de aprendizaje tales como el aprendizaje por refuerzo y aprendizaje profundo.

Contenido

1. Introducción.
2. Adquisición, Procesamiento y Etiquetado de Datos.
3. Extracción y Selección de Características.
4. Aprendizaje Supervisado.
5. Aprendizaje NO Supervisado.
6. Técnicas Modernas de Aprendizaje de Máquinas.
7. Aplicaciones y Casos de Éxito.

APRENDIZAJE DE MÁQUINAS
Introducción

JOHN W. BRANCH

Profesor Titular
Departamento de Ciencias de la Computación y de la Decisión
Director del Grupo de I+D en Inteligencia Artificial – GIDIA
jwbranch@unal.edu.co

Gidia Grupo de I+D en Inteligencia Artificial

UNIVERSIDAD NACIONAL DE COLOMBIA

Contenido

1. Definición de aprendizaje de máquinas.
2. Aplicaciones clásicas del aprendizaje de máquinas.
3. Tipos de aprendizaje:
   - Supervisado.
   - No Supervisado.
4. Metodología clásica para el desarrollo de aplicaciones de aprendizaje de máquinas.
5. Herramientas Tecnológicas.
6. Aplicaciones.

Motivación

OBSERVE EL VIDEO Y RESPONDA A LAS SIGUIENTES PREGUNTAS:

¿Cuántos datos se requieren para entrenar un sistema de visión artificial?

¿Es posible decir que los computadores ya sobrepasaron la capacidad humana?

¿Qué problemas evidencian los sistemas de visión artificial, y en general de los sistemas de Reconocimiento de Patrones?

https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures?language=es

Gidia Grupo de I+D en Inteligencia Artificial

UNIVERSIDAD NACIONAL DE COLOMBIA

Aprendizaje de Máquinas y Predicción

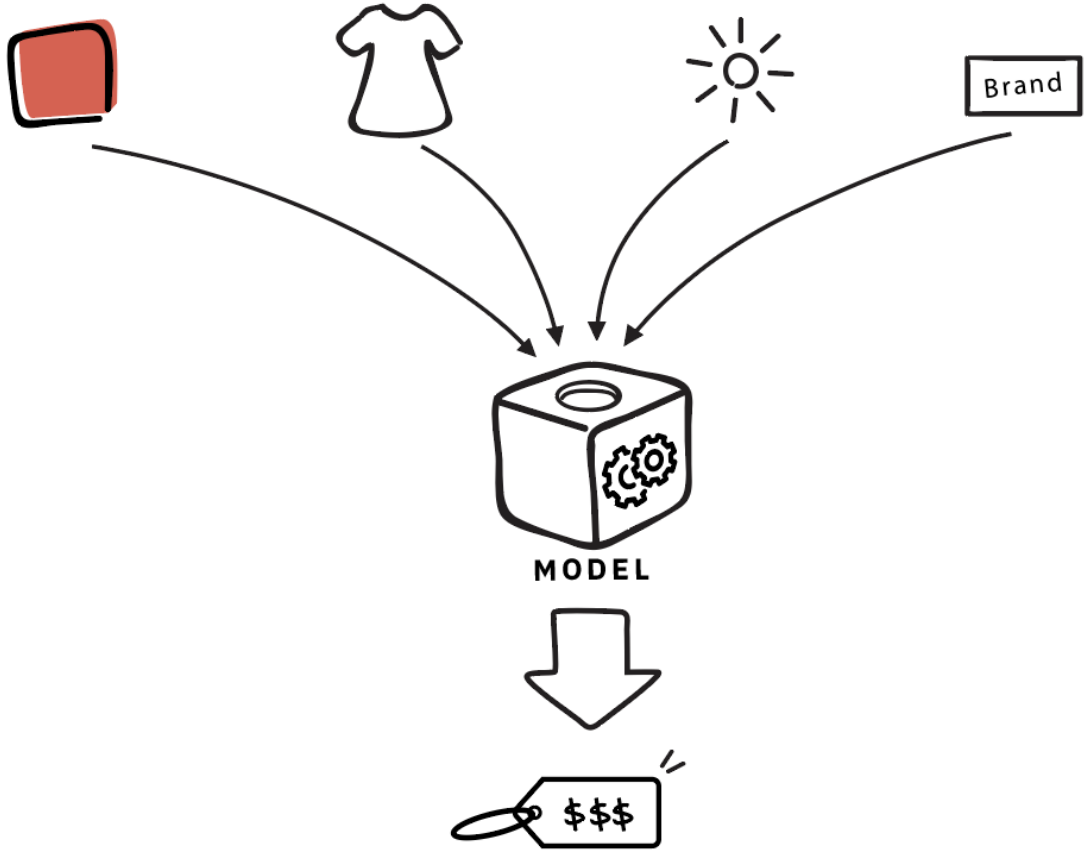
- El problema central del aprendizaje de máquinas es la ***predicción***, es decir, aplicar sobre datos nuevos un algoritmo que ha sido entrenado sobre un conjunto de datos históricos.

- Aunque suene como predecir el futuro, el término *predicción* generalmente se usa para el procesamiento de datos nuevos *in-situ*. Cuando los datos tienen un componente temporal se utiliza el término *pronóstico*.

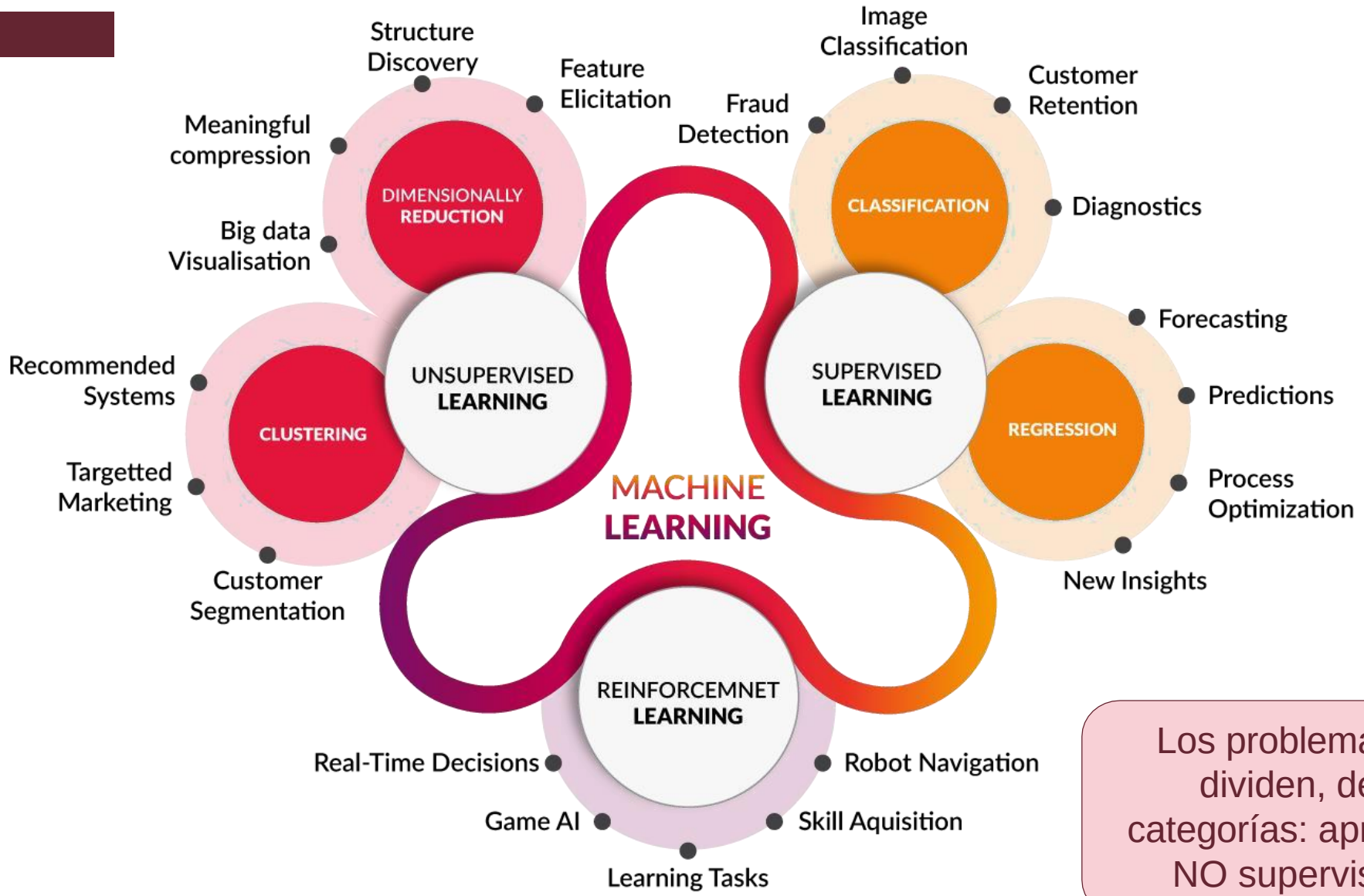
- En este orden de ideas, cuando se habla de predicción se puede hacer referencia a:

  - ***Clasificación*** para obtener una etiqueta o clase conocida.
  - ***Regresión*** para obtener un valor numérico.

  - ***Agrupamiento*** para descubrir etiquetas o patrones nuevos a partir de datos no etiquetados a partir de



Tomada de: https://medium.com/@srnghn



Los problemas de aprendizaje de máquina se dividen, de manera muy general, en tres categorías: aprendizaje supervisado, aprendizaje NO supervisado y aprendizaje por refuerzo.

Tomada de: https://www.predictiveanalyticsworld.com/patimes/wp-content/uploads/2018/01/machinelearning-IMAGE.png

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

Data is not labeled in any way

SUPERVISED

Predict a category

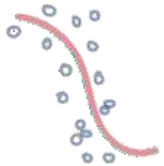
Predict a number

CLASSIFICATION

«Divide the socks by color»

REGRESSION

«Divide the ties by length»



UNSUPERVISED

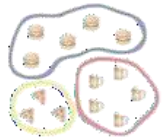
Divide by similarity

Find hidden dependencies

Identify sequences

CLUSTERING

«Split up similar clothing into stacks»

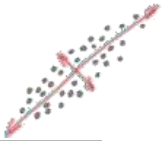


ASSOCIATION

«Find what clothes I often wear together»

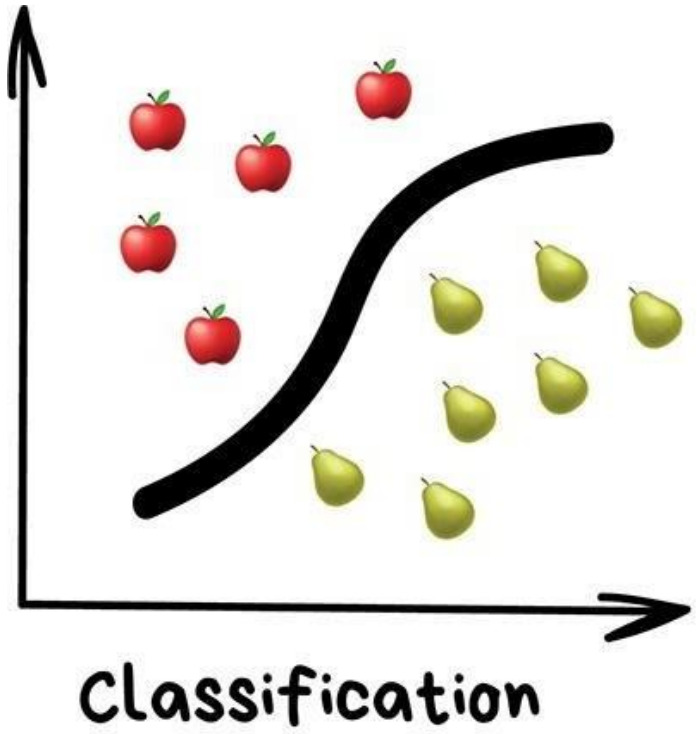
DIMENSION REDUCTION (generalization)

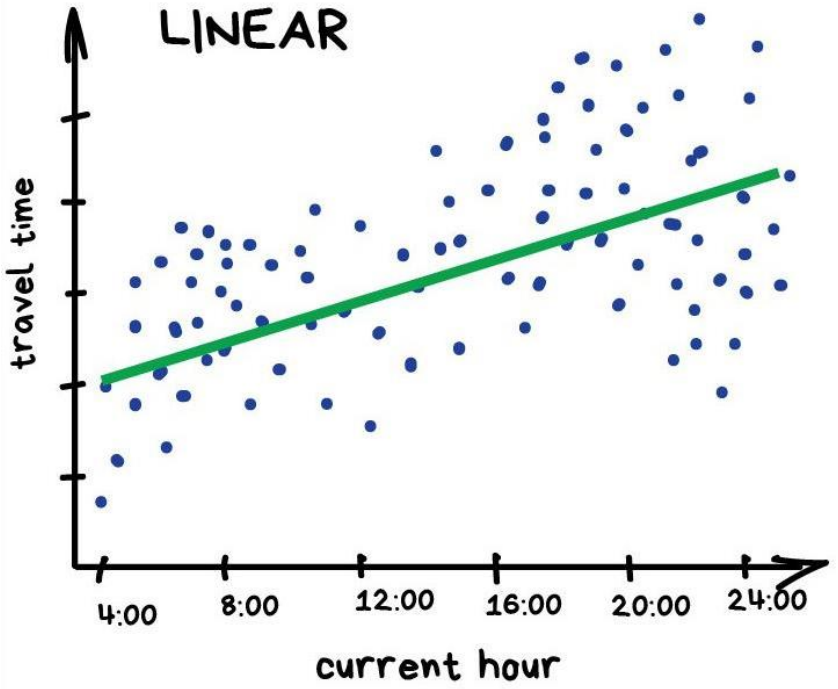
«Make the best outfits from the given clothes»

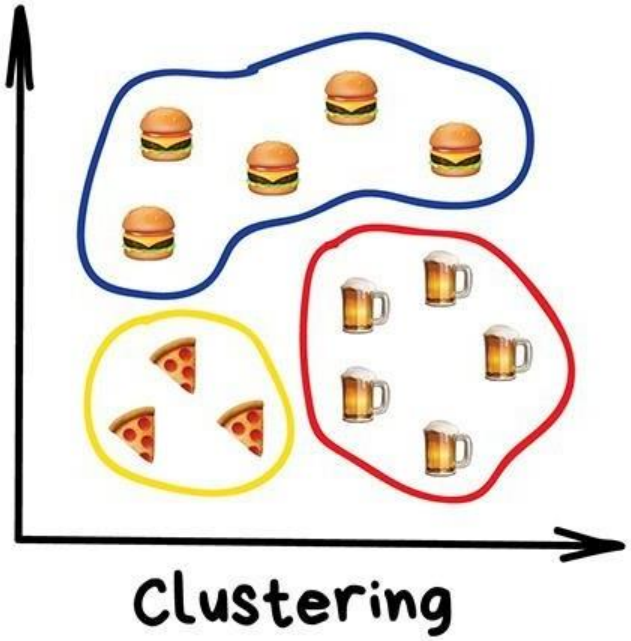


Tomada de: https://vas3k.com/blog/machine_learning/

Tipos de Aprendizaje







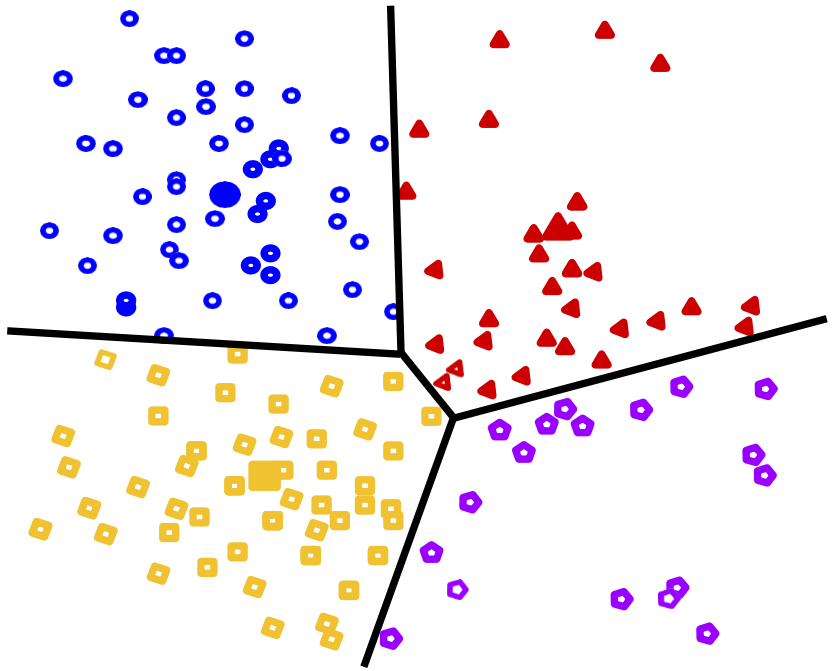
Aprendizaje Supervisado vs Aprendizaje NO Supervisado

Aprendizaje Supervisado: requiere de un conjunto de datos conocidos a partir del cual se crea un **modelo** para predecir el valor de una variable de salida. El aprendizaje supervisado se puede usar en dos tareas:

- **Clasificación:** en este caso la variable de salida es una etiqueta que determina la clase a la que pertenecen los datos de entrada, es decir, la variable de salida es una **variable discreta**.
- **Regresión:** en este caso los algoritmos de aprendizaje buscan predecir el valor de una **variable continua** a partir de los datos de entrada. Un ejemplo de una tarea de regresión es el de estimar la longitud de un salmón en función de su edad y su peso.

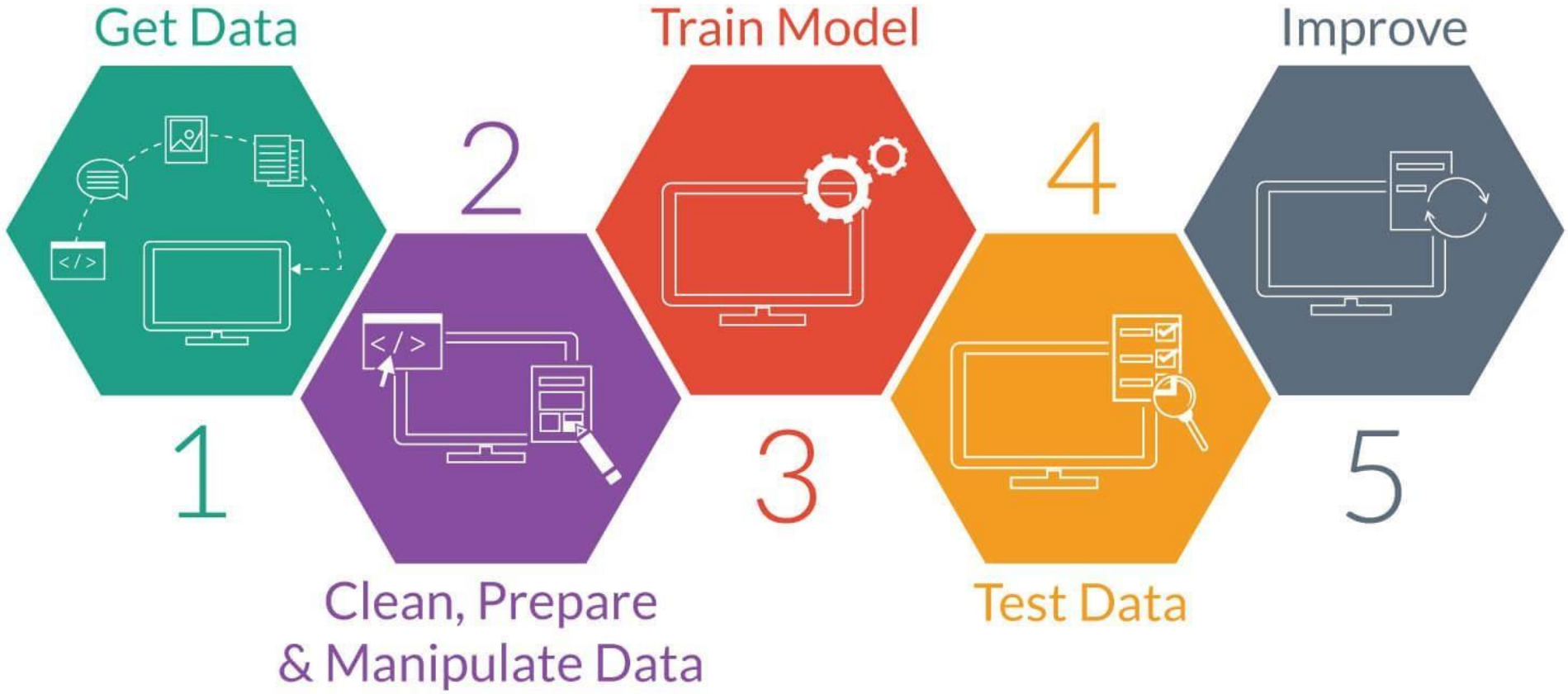
Aprendizaje Supervisado vs Aprendizaje NO Supervisado

Aprendizaje No Supervisado: se cuenta con un conjunto de datos de entrenamiento, pero no hay una variable específica de salida (se desconocen las clases). En este sentido, el objetivo de los problemas del aprendizaje no supervisado es, por ejemplo, el de agrupar los datos de entrada con base en algún criterio de similitud o disimilitud o determinar la distribución estadística de los datos, conocida como estimación de la densidad.



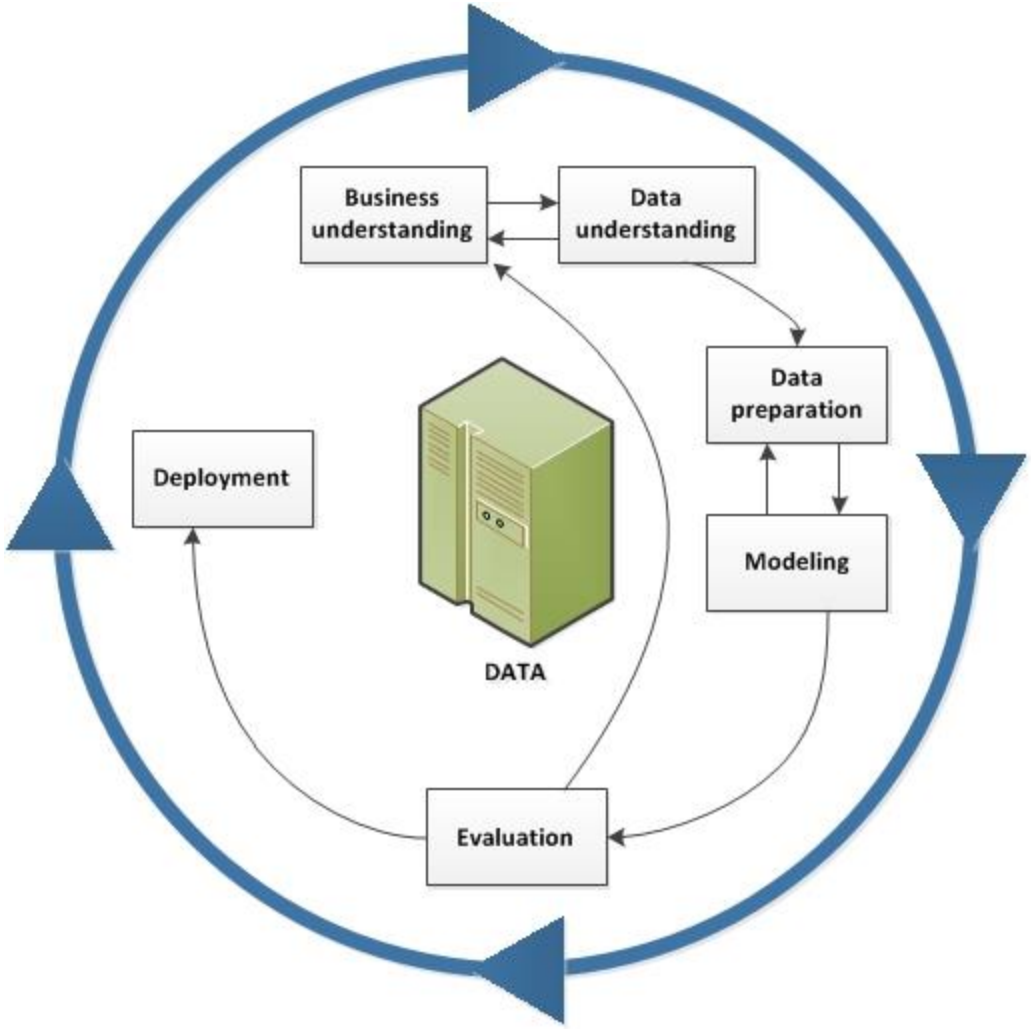
Metodología Clásica para el Desarrollo

https://www.portalveterinaria.com/porcino/actualidad/31080/em-pig-data-em.html



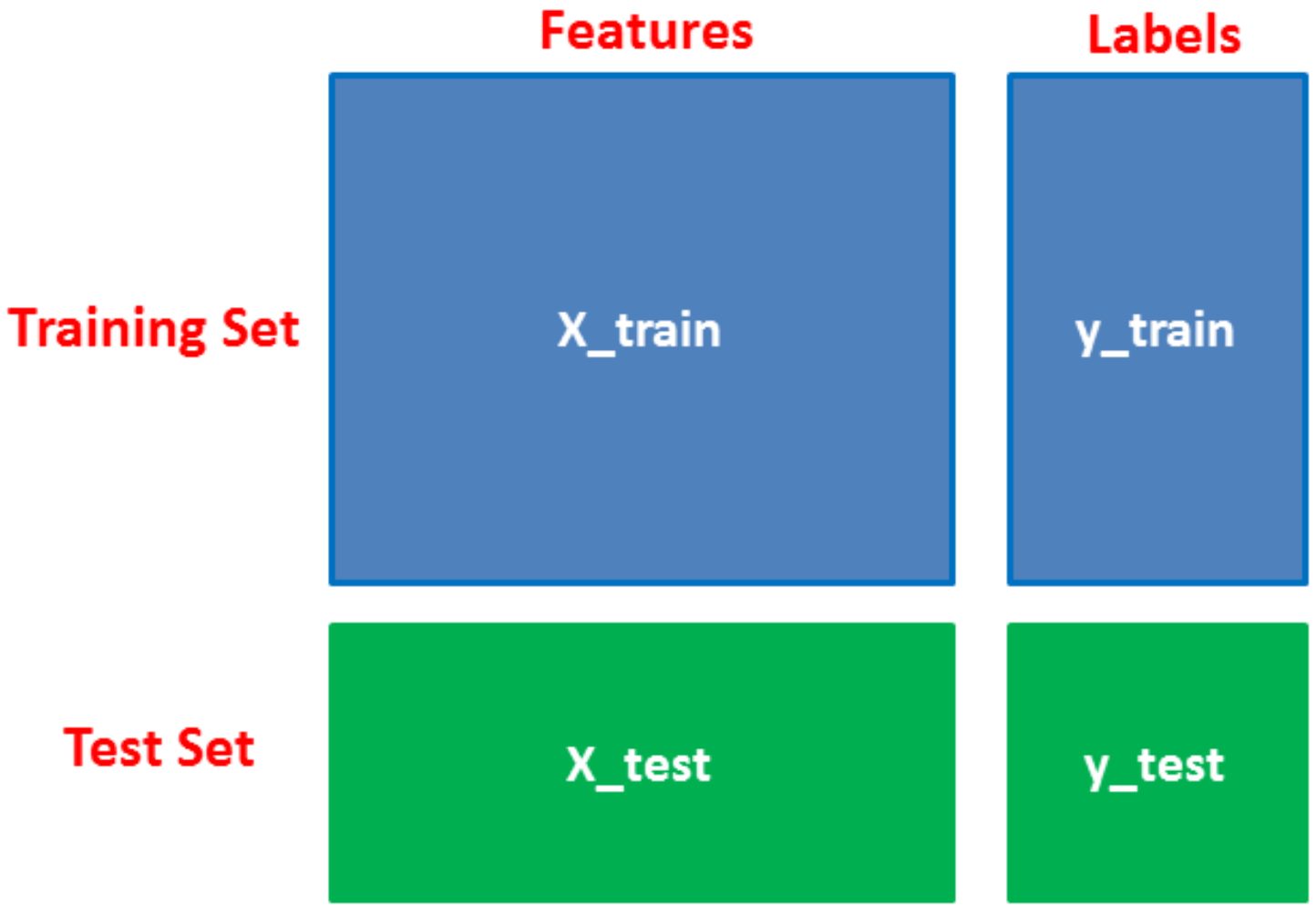
Tomada de: https://newtiummedia.blob.core.windows.net/images/Steps-to-Predictive-Modelling.jpg

Metodología iterativa: CRISP-DM



Tomada de: https://www.ibm.com/

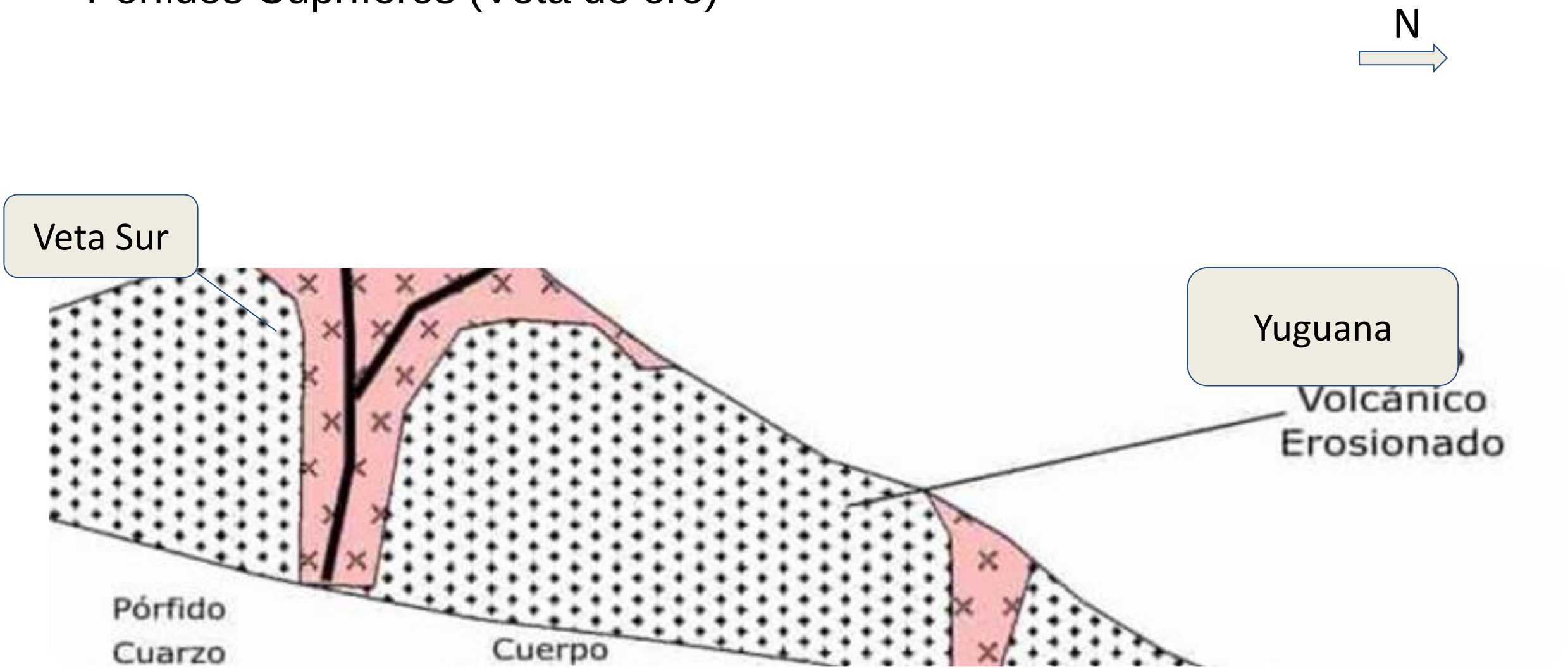
Partición del Conjunto de Datos



EJEMPLO

Predicción en la recuperación del oro con técnicas de aprendizaje de máquina

Pórfidos Cupríferos (Veta de oro)



Proceso de producción del Oro

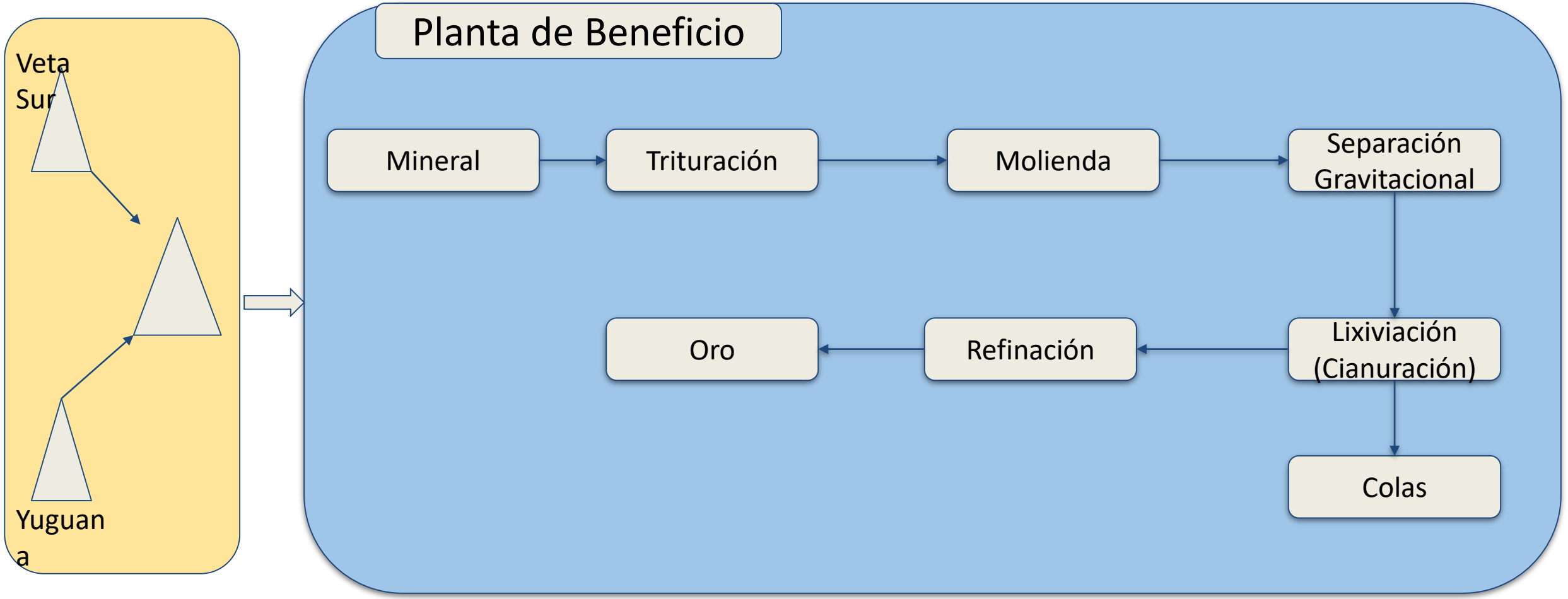
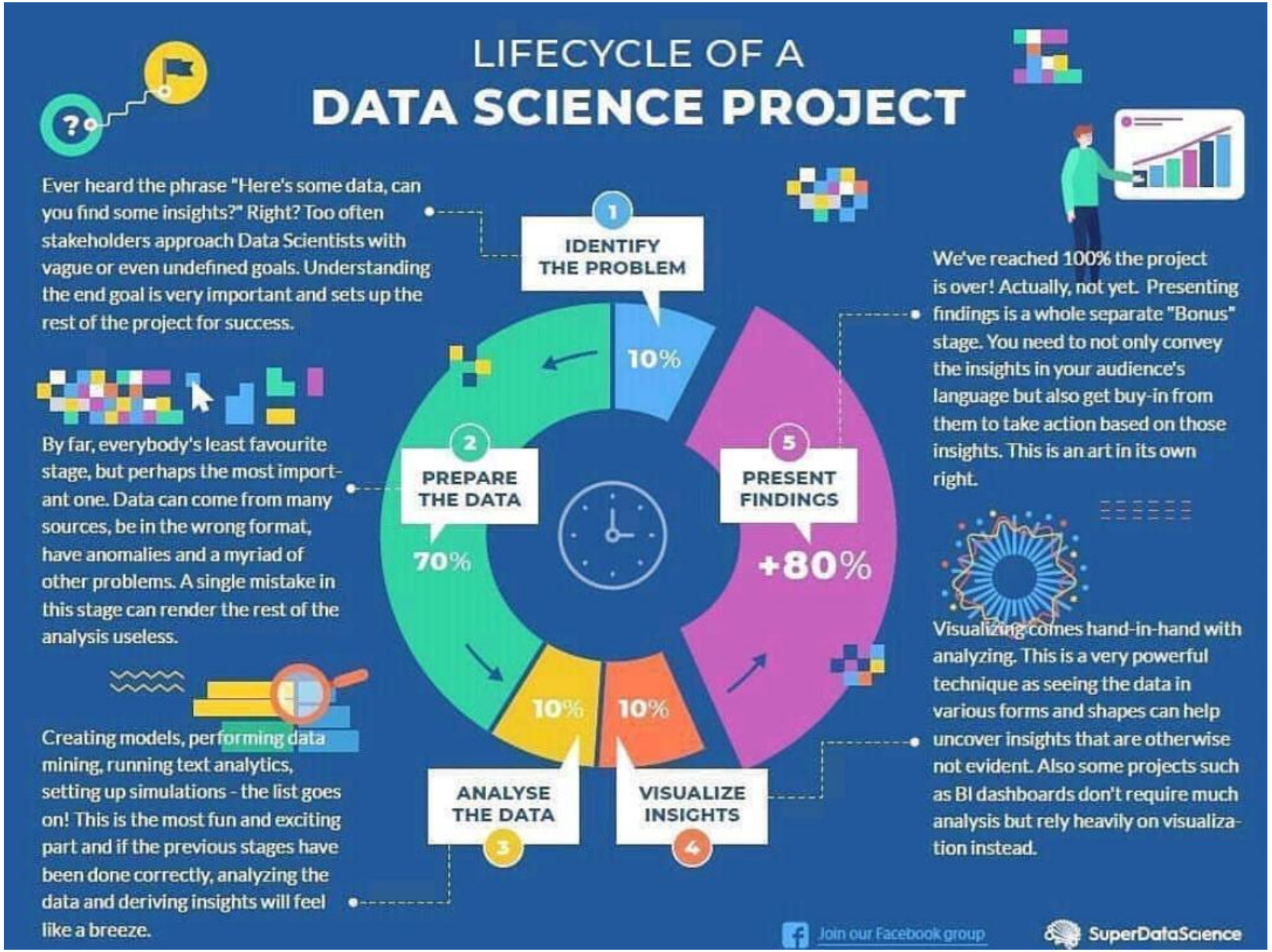
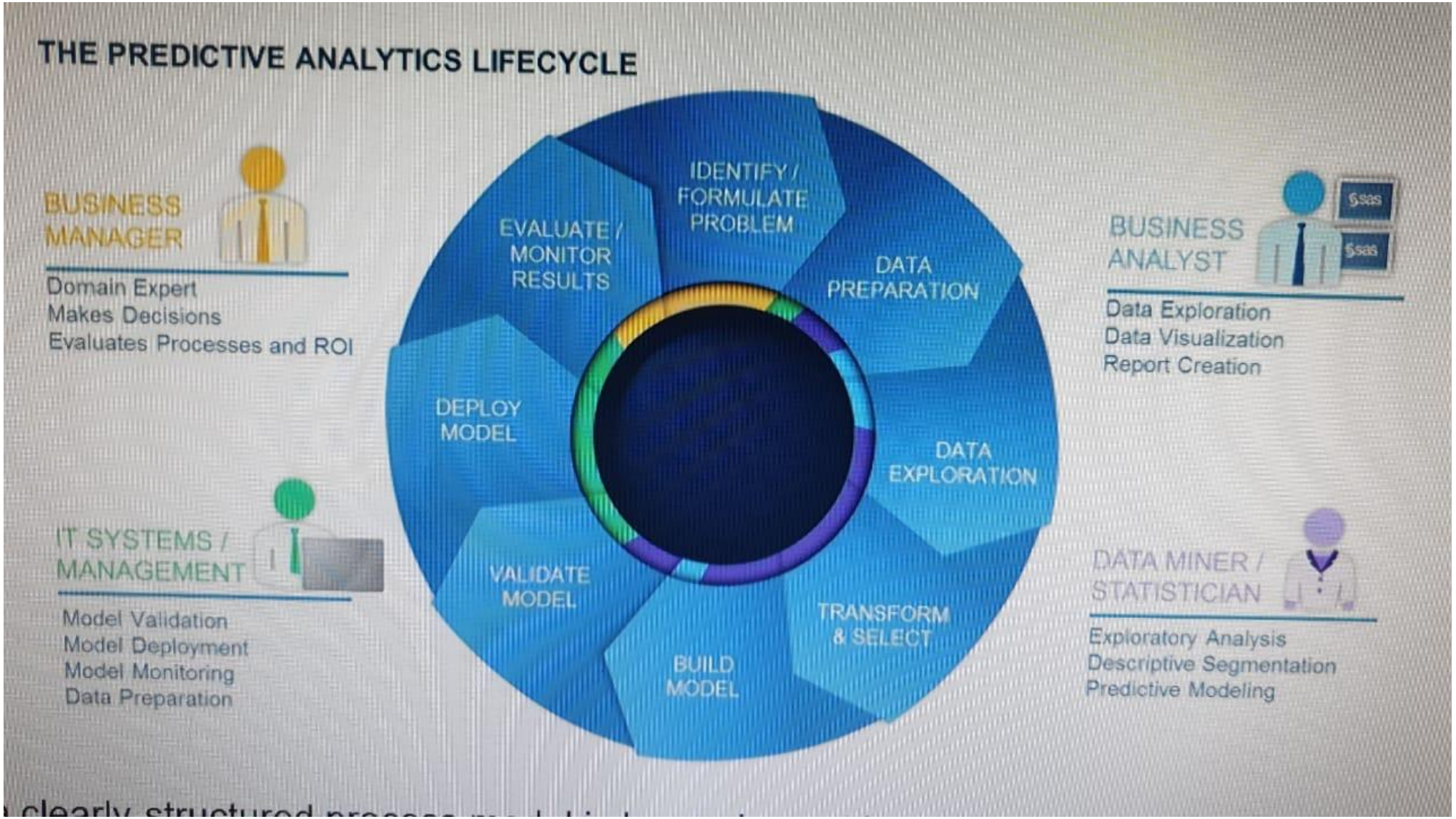


Tabla de Datos

	Composite	# test	Au	Ag	Cu	Pb	Zn	Fe	As	Au			Ag		
										% Rec. Gravity	% Rec. Leach	Total	% Rec. Gravity	% Rec. Leach	Total
3	VSYU-0	17	19,987	44,534	779,020	1264,985	5458,144	8,424	265,511	73,815	22,088	95,903	11,767	49,936	61,703
4	VSYU-0	22	20,151	49,273	632,720	1264,985	5458,144	8,424	265,511	73,211	22,428	95,639	10,635	68,045	78,680
5	VSYU-0	25	19,968	45,890	674,844	1264,985	5458,144	8,424	265,511	73,885	22,015	95,899	11,419	48,500	59,919
6	VSYU-0	39	19,458	41,018	670,635	1264,985	5458,144	8,424	265,511	75,822	21,407	97,229	12,776	44,281	57,057
7	VSYU-0	55	20,561	43,457	675,558	1264,985	5458,144	8,424	265,511	71,751	24,752	96,503	12,059	56,232	68,291
8	VSYU-0	61	19,509	46,377	705,536	1435,872	4441,843	6,469	265,511	75,621	21,922	97,543	11,300	60,193	71,493
9	VSYU-0	67	20,618	47,777	740,652	1625,530	4414,298	6,308	265,511	71,554	24,862	96,416	10,968	63,701	74,669
10	VSYU-0	73	20,216	47,312	663,898	1254,999	4932,907	5,741	265,511	72,979	24,057	97,036	11,076	60,474	71,550
11	VSYU-0	80	21,252	51,169	1007,623	1102,351	2244,770	8,590	265,500	68,880	28,677	97,557	10,162	68,138	78,299
12	VSYU-0	81	20,839	45,506	697,557	1313,914	4822,066	6,972	265,511	70,794	25,972	96,766	11,516	55,405	66,921
13	VSYU-0	91	20,293	42,891	728,418	1303,626	6610,936	7,907	265,511	72,699	24,053	96,752	12,218	53,279	65,497
14	VSYO-20	15	20,753	44,773	836,413	1141,886	2460,446	9,965	474,271	68,982	25,819	94,801	8,737	54,846	63,583
15	VSYO-20	16	21,732	47,582	818,915	1141,886	2460,446	9,965	474,271	65,876	29,159	95,035	8,221	52,263	60,484
16	VSYO-20	23	21,716	51,942	800,256	1141,886	2460,446	9,965	474,271	65,924	29,016	94,939	7,531	70,041	77,572


LIFECYCLE OF A
DATA SCIENCE PROJECT
Ever heard the phrase "Here's some data, can you find some insights?" Right? Too often stakeholders approach Data Scientists with vague or even undefined goals. Understanding the end goal is very important and sets up the rest of the project for success.
1
IDENTIFY THE PROBLEM
10%
By far, everybody's least favourite stage, but perhaps the most important one. Data can come from many sources, be in the wrong format, have anomalies and a myriad of other problems. A single mistake in this stage can render the rest of the analysis useless.
2
PREPARE THE DATA
70%
Creating models, performing data mining, running text analytics, setting up simulations - the list goes on! This is the most fun and exciting part and if the previous stages have been done correctly, analyzing the data and deriving insights will feel like a breeze.
3
ANALYSE THE DATA
10%
4
VISUALIZE INSIGHTS
10%
5
PRESENT FINDINGS
+80%
We've reached 100% the project is over! Actually, not yet. Presenting findings is a whole separate "Bonus" stage. You need to not only convey the insights in your audience's language but also get buy-in from them to take action based on those insights. This is an art in its own right.
Visualizing comes hand-in-hand with analyzing. This is a very powerful technique as seeing the data in various forms and shapes can help uncover insights that are otherwise not evident. Also some projects such as BI dashboards don't require much analysis but rely heavily on visualization instead.
Join our Facebook group
SuperDataScience


THE PREDICTIVE ANALYTICS LIFECYCLE
BUSINESS MANAGER
Domain Expert
Makes Decisions
Evaluates Processes and ROI
IT SYSTEMS / MANAGEMENT
Model Validation
Model Deployment
Model Monitoring
Data Preparation
IDENTIFY / FORMULATE PROBLEM
DATA PREPARATION
DATA EXPLORATION
TRANSFORM & SELECT
BUILD MODEL
VALIDATE MODEL
DEPLOY MODEL
EVALUATE / MONITOR RESULTS
BUSINESS ANALYST
Data Exploration
Data Visualization
Report Creation
DATA MINER / STATISTICIAN
Exploratory Analysis
Descriptive Segmentation
Predictive Modeling

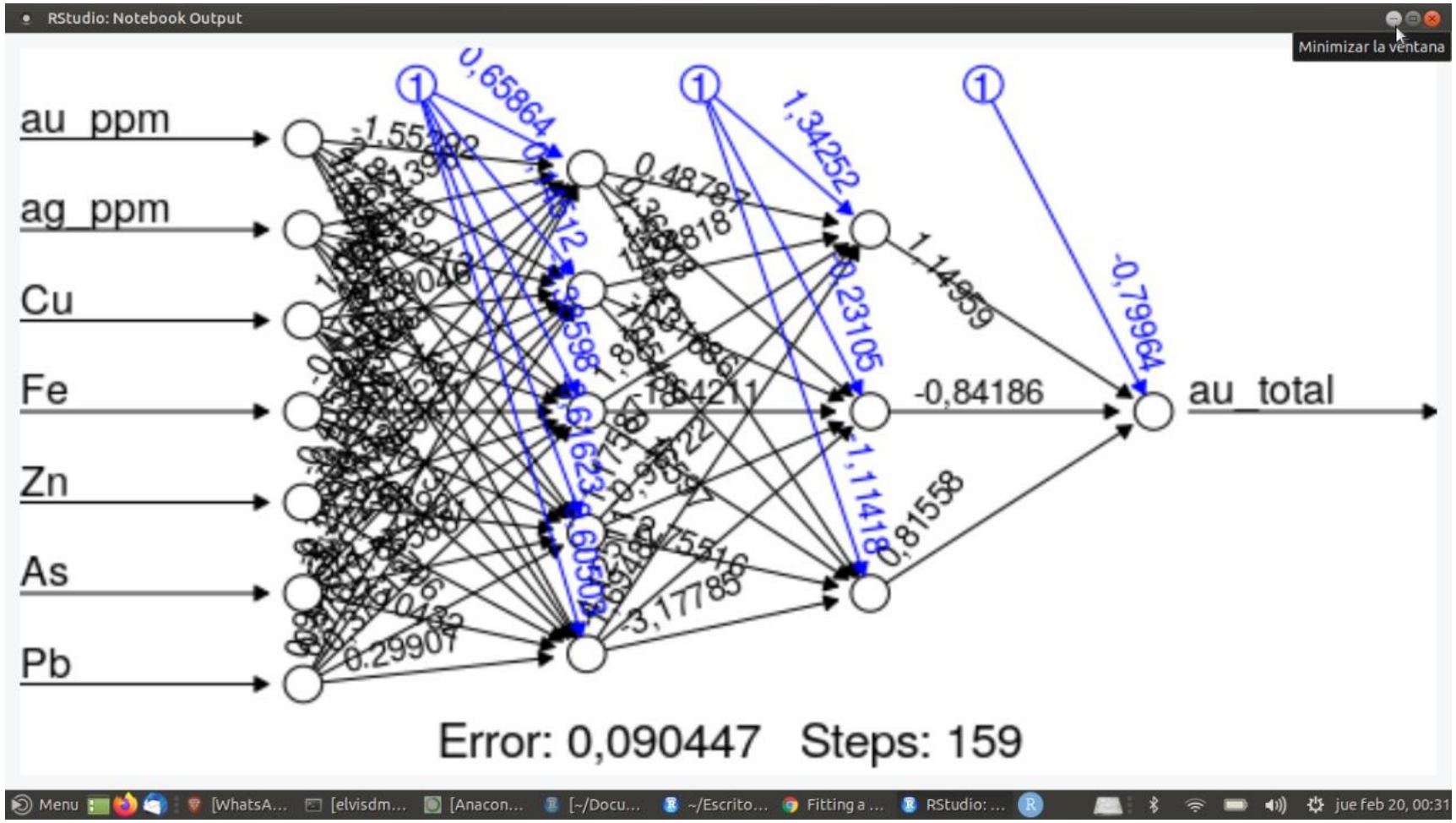
Problema

¿Se puede predecir cuánto se va a recuperar de oro al final del proceso minero?

Datos preparados

	test	au_ppm	ag_ppm	Cu	Pb	Zn	Fe	As	au_gravity	au_leach	au_total	ag_gravity	ag_leach	ag_total	composite
1	22	20	49	632	1264	5458	8	265	73.21110	22.42830	95.63940	10.635488	68.04496	78.68045	VSYU-0
2	25	19	45	674	1264	5458	8	265	73.88466	22.01467	95.89933	11.419418	48.49958	59.91900	VSYU-0
3	39	19	41	670	1264	5458	8	265	75.82190	21.40685	97.22875	12.775838	44.28111	57.05695	VSYU-0
4	55	20	43	675	1264	5458	8	265	71.75123	24.75215	96.50338	12.058712	56.23211	68.29082	VSYU-0
5	61	19	46	705	1435	4441	6	265	75.62149	21.92169	97.54318	11.299658	60.19290	71.49256	VSYU-0
6	67	20	47	740	1625	4414	6	265	71.55423	24.86189	96.41612	10.968357	63.70083	74.66919	VSYU-0
7	73	20	47	663	1254	4932	5	265	72.97881	24.05749	97.03630	11.076197	60.47359	71.54979	VSYU-0
8	80	21	51	1007	1102	2244	8	265	68.88001	28.67667	97.55668	10.161628	68.13784	78.29947	VSYU-0
9	81	20	45	697	1313	4822	6	265	70.79412	25.97152	96.76564	11.515846	55.40471	66.92056	VSYU-0
10	91	20	42	728	1303	6610	7	265	72.69893	24.05350	96.75243	12.218081	53.27897	65.49706	VSYU-0
11	15	20	44	836	1141	2460	9	474	68.98168	25.81912	94.80079	8.736721	54.84646	63.58318	VSYO-20
12	16	21	47	818	1141	2460	9	474	65.87594	29.15894	95.03488	8.220964	52.26279	60.48375	VSYO-20
13	23	21	51	800	1141	2460	9	474	65.92357	29.01570	94.93927	7.530982	70.04138	77.57236	VSYO-20
14	40	21	44	752	1141	2460	9	474	68.03405	26.93313	94.96718	8.788163	48.78963	57.57779	VSYO-20
15	71	18	45	764	1170	4093	6	474	78.16934	17.57556	95.74489	8.593709	66.77955	75.37326	VSYO-20
16	72	21	48	778	1238	4141	5	474	66.29339	31.85603	98.14941	8.097376	66.17509	74.27247	VSYO-20

Modelo de Predicción



Gidia Grupo de I+D en Inteligencia Artificial

UNIVERSIDAD NACIONAL DE COLOMBIA

Predichos vs Observados

Error = 2,59

prediccion <dbl>	real <dbl>
95.78129	97.54318
95.64481	94.96718
94.51549	95.74489
96.42540	98.14941
95.95849	95.72399
93.77977	94.49732
95.12479	92.86778
93.87093	95.19341
92.19314	92.64212
89.73256	94.58457

Herramientas Tecnológicas

Tomada de: https://dataflog.com/read/machine-learning-becomes-mainstream-advantage/2236

Herramientas Tecnológicas

Preguntas

EVALUACIÓN

Seguimiento	40%
Proyecto Final del Curso	60%

Reconocimiento de Patrones

Ejemplo Práctico - Mandarinas vs. Naranjas

[Capítulo 1]

Domingo Mery

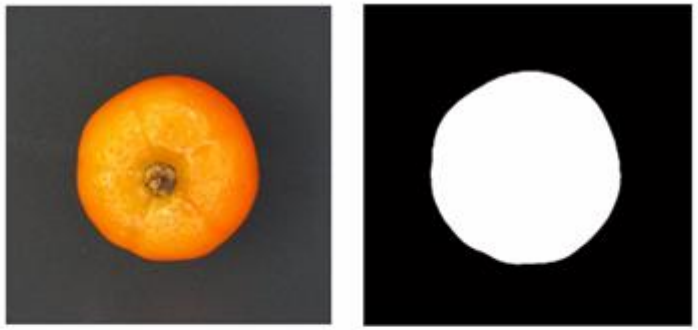
Departmento de Ciencia de la Computación
Escuela de Ingeniería
Universidad Católica de Chile

Un ejemplo práctico

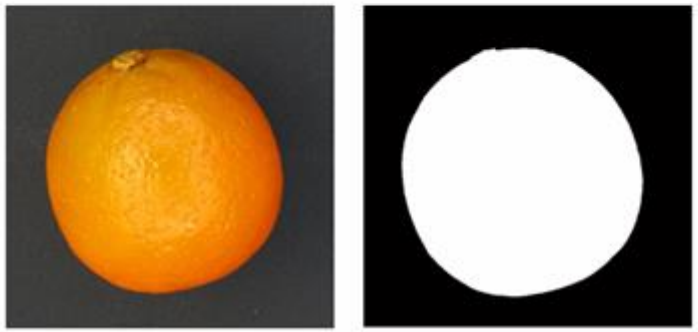
¿cómo separar las mandarinas de las naranjas?

Un ejemplo práctico

Medición del tamaño es una buena alternativa: (*las mandarinas son más pequeñas*)



Área = 15.457 pixeles



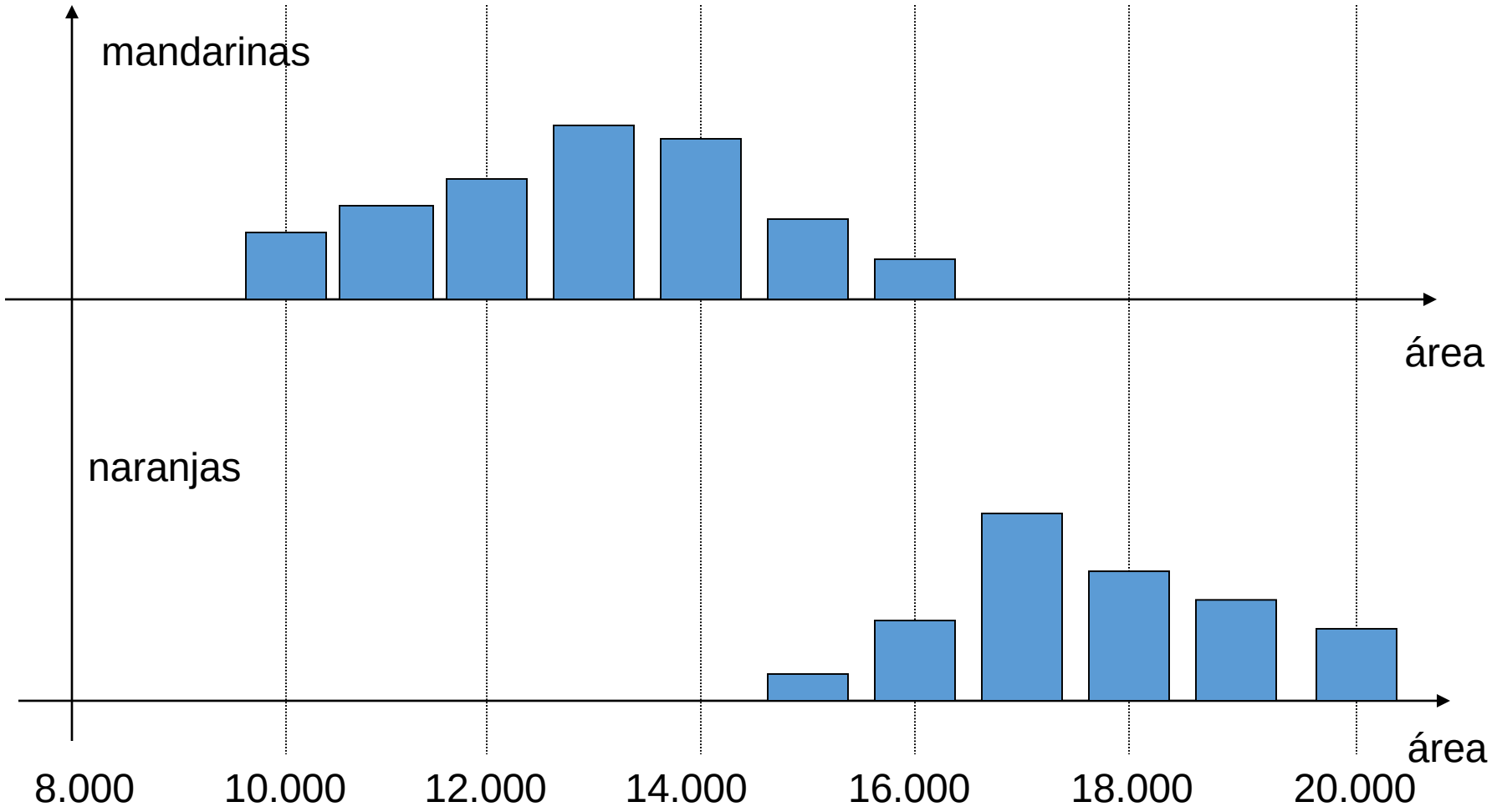
Área = 18.583 pixeles

Recolección de información: Área en Pixeles

Naranja-01	19.327	Mandarina-01	13.221
Naranja-02	18.265	Mandarina-02	14.987
Naranja-03	17.456	Mandarina-03	15.321
Naranja-04	19.341	Mandarina-04	15.987
Naranja-05	16.342	Mandarina-05	16.345
Naranja-06	16.987	Mandarina-06	15.965
Naranja-07	17.001	Mandarina-07	16.341
:	19.056	:	
Naranja-75	15.900	Mandarina-50	13.439

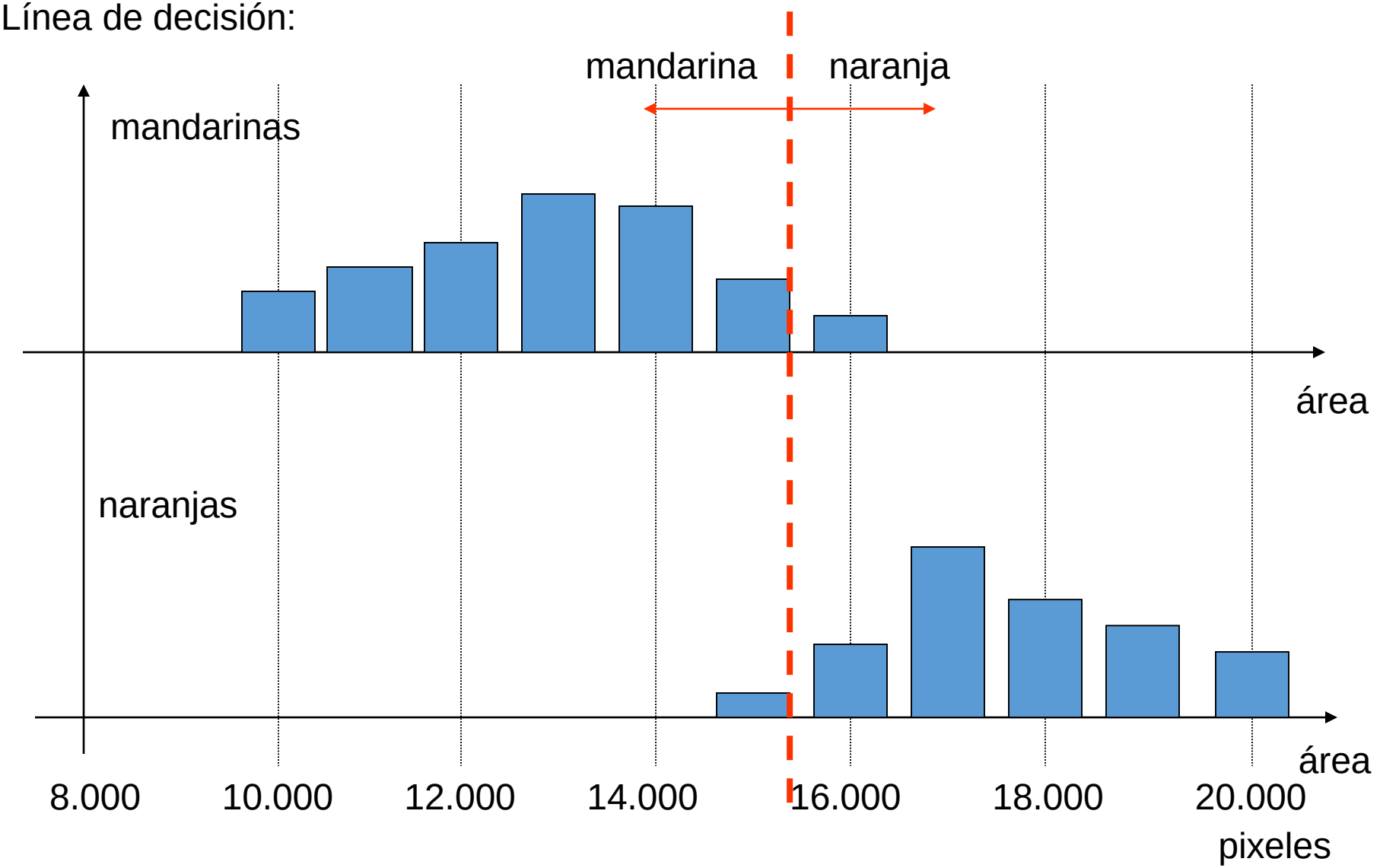
Un ejemplo práctico

Histogramas:

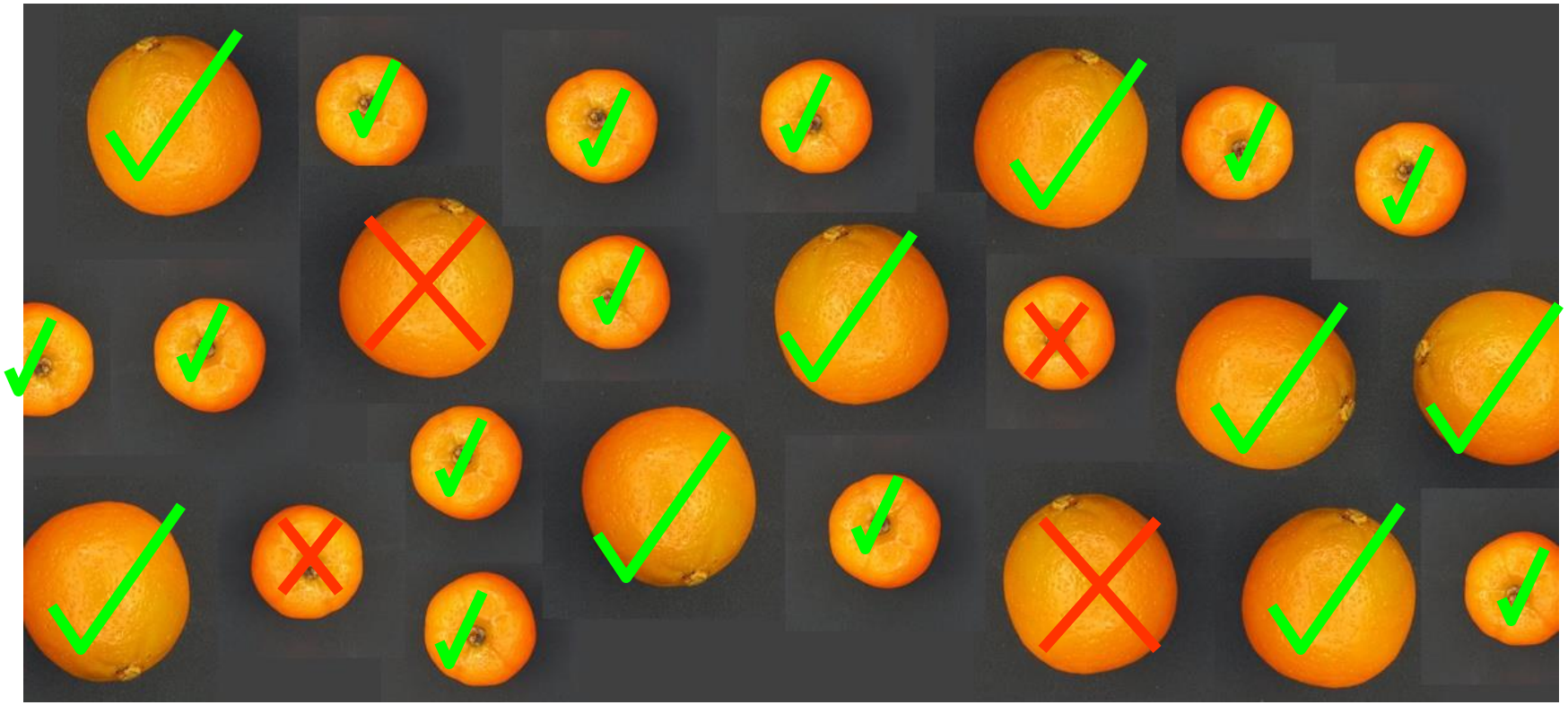


Un ejemplo práctico

Línea de decisión:

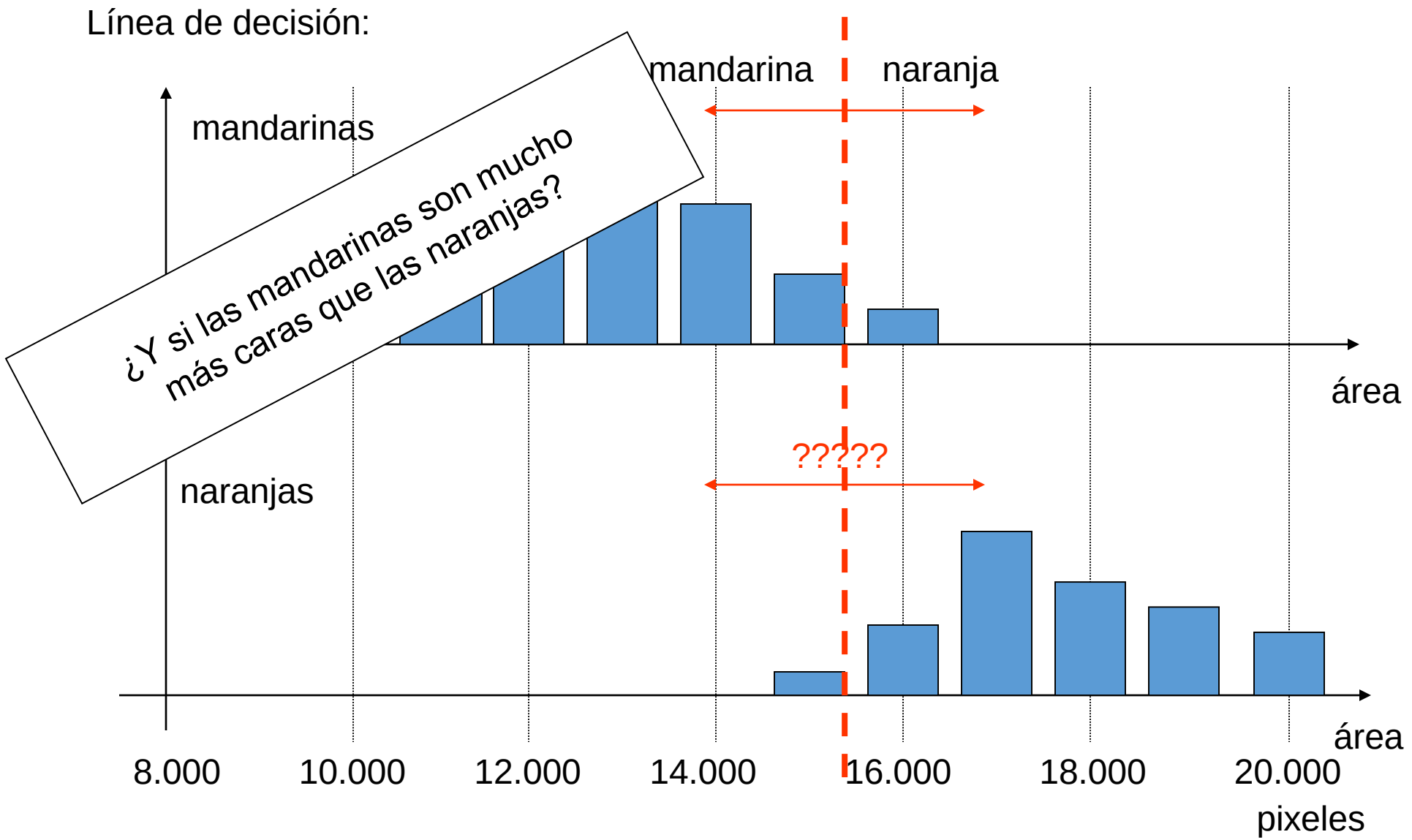


Un ejemplo práctico

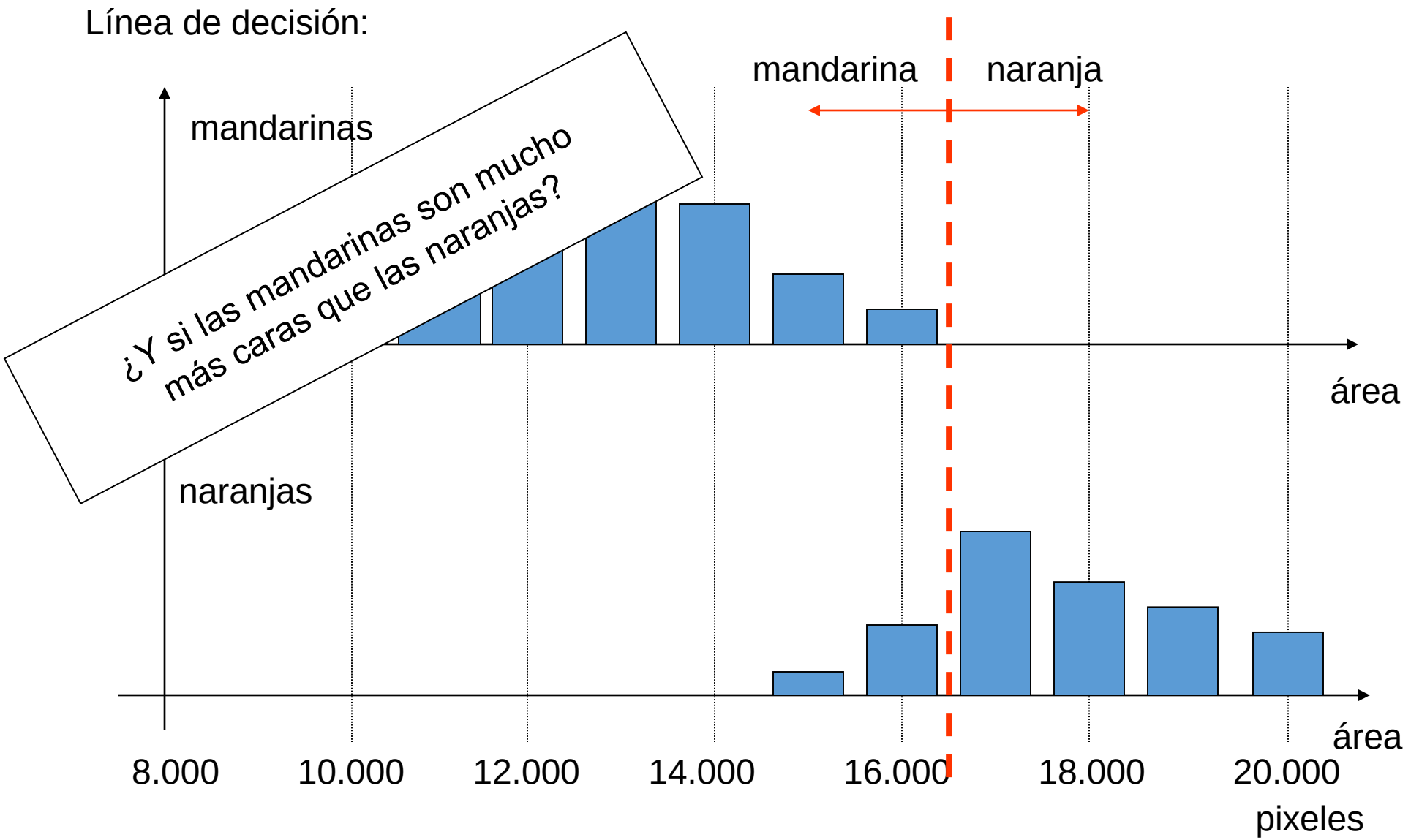


La separación es buena pero no es perfecta

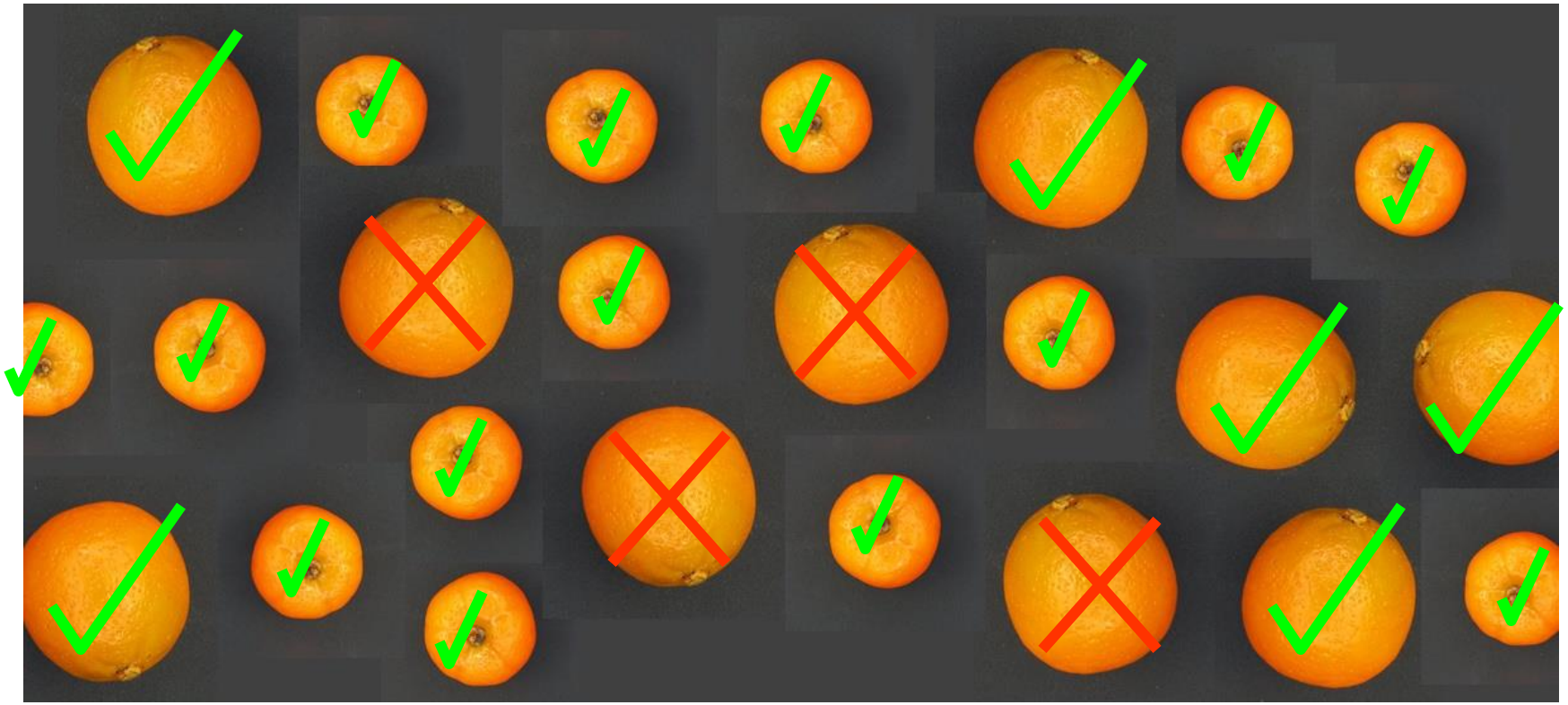
Un ejemplo práctico



Un ejemplo práctico



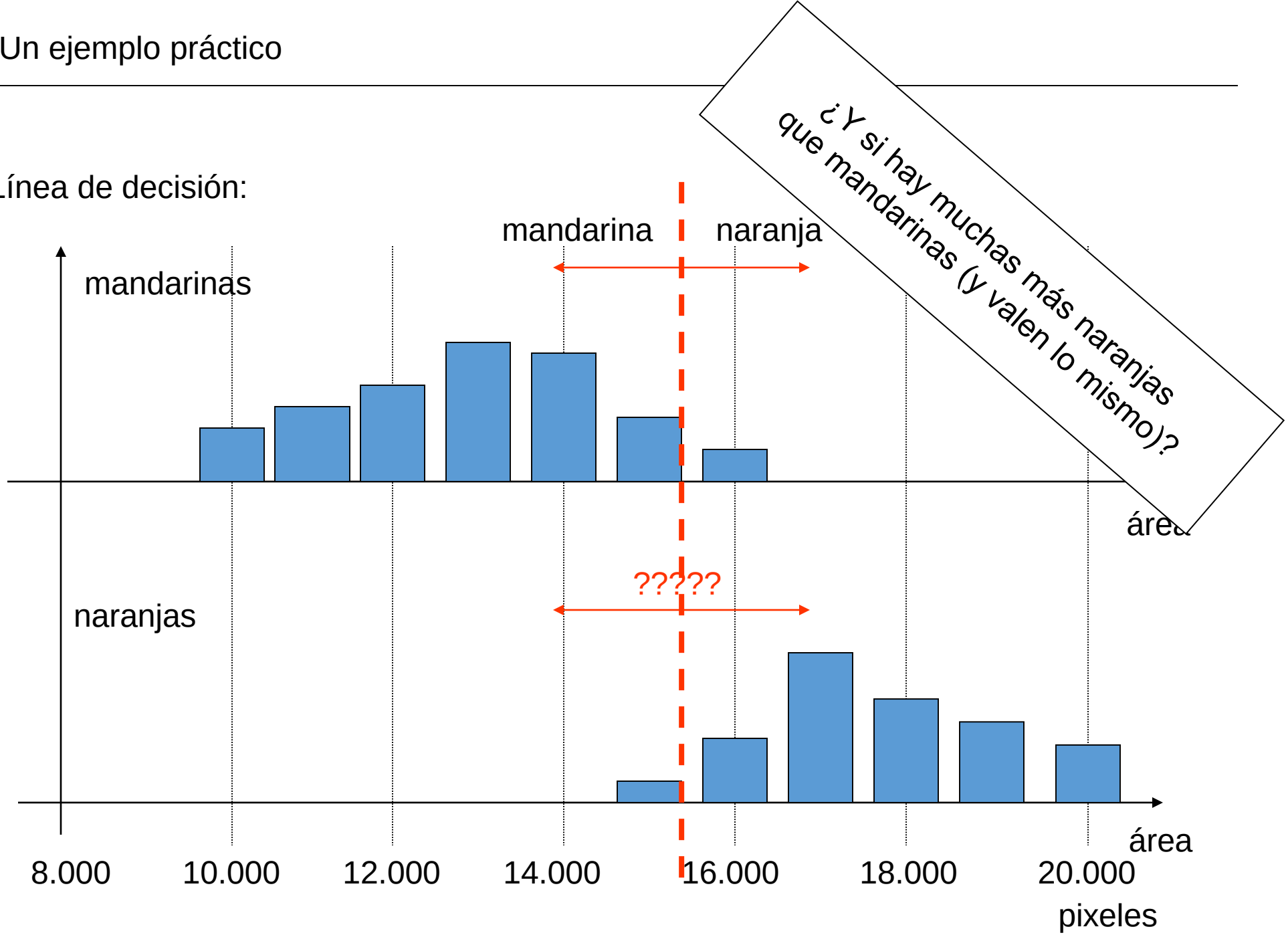
Un ejemplo práctico



Todas las mandarinas son clasificadas perfectamente
… pero el costo es que hay varias naranjas mal clasificadas

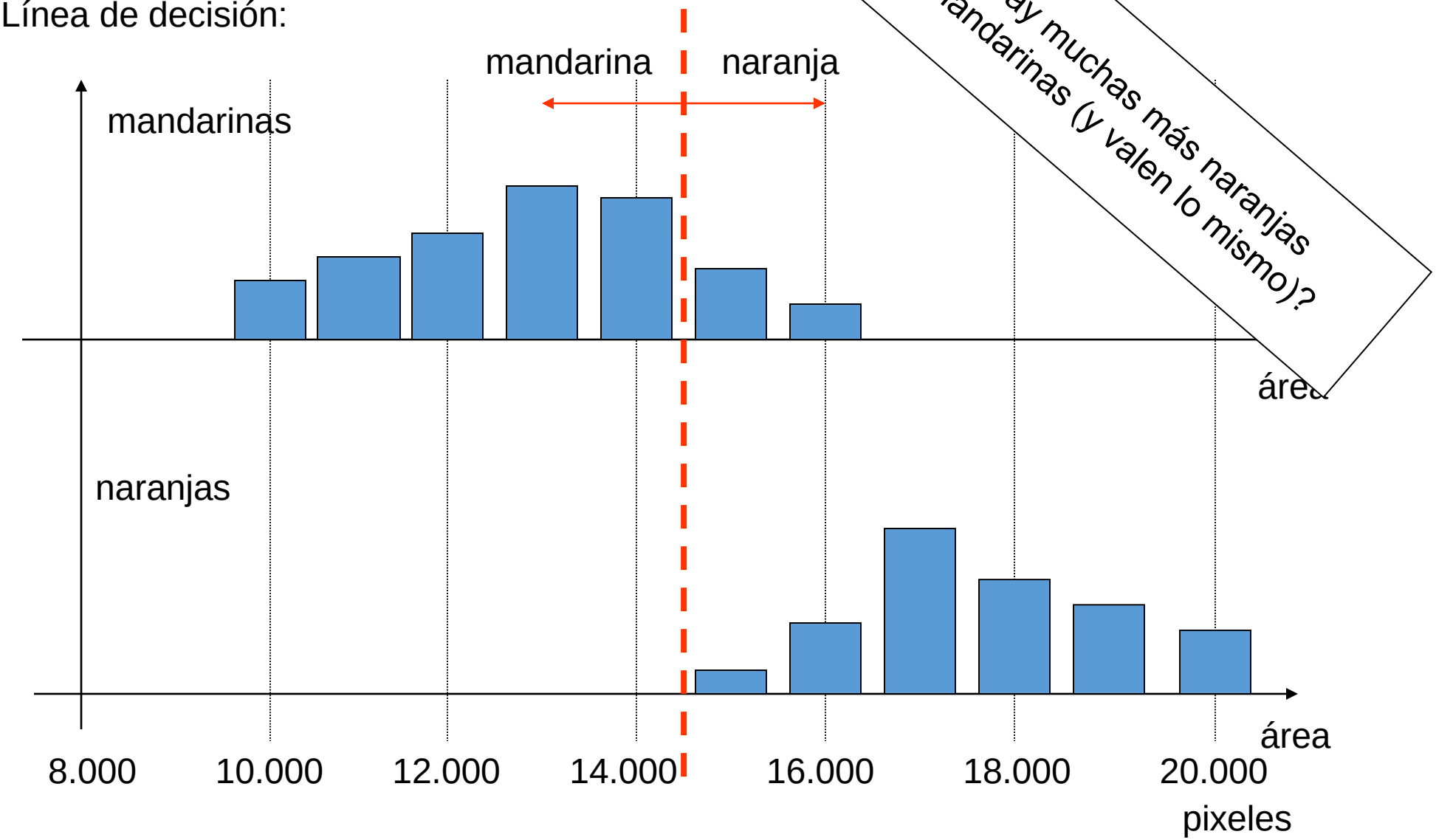
Un ejemplo práctico

Línea de decisión:



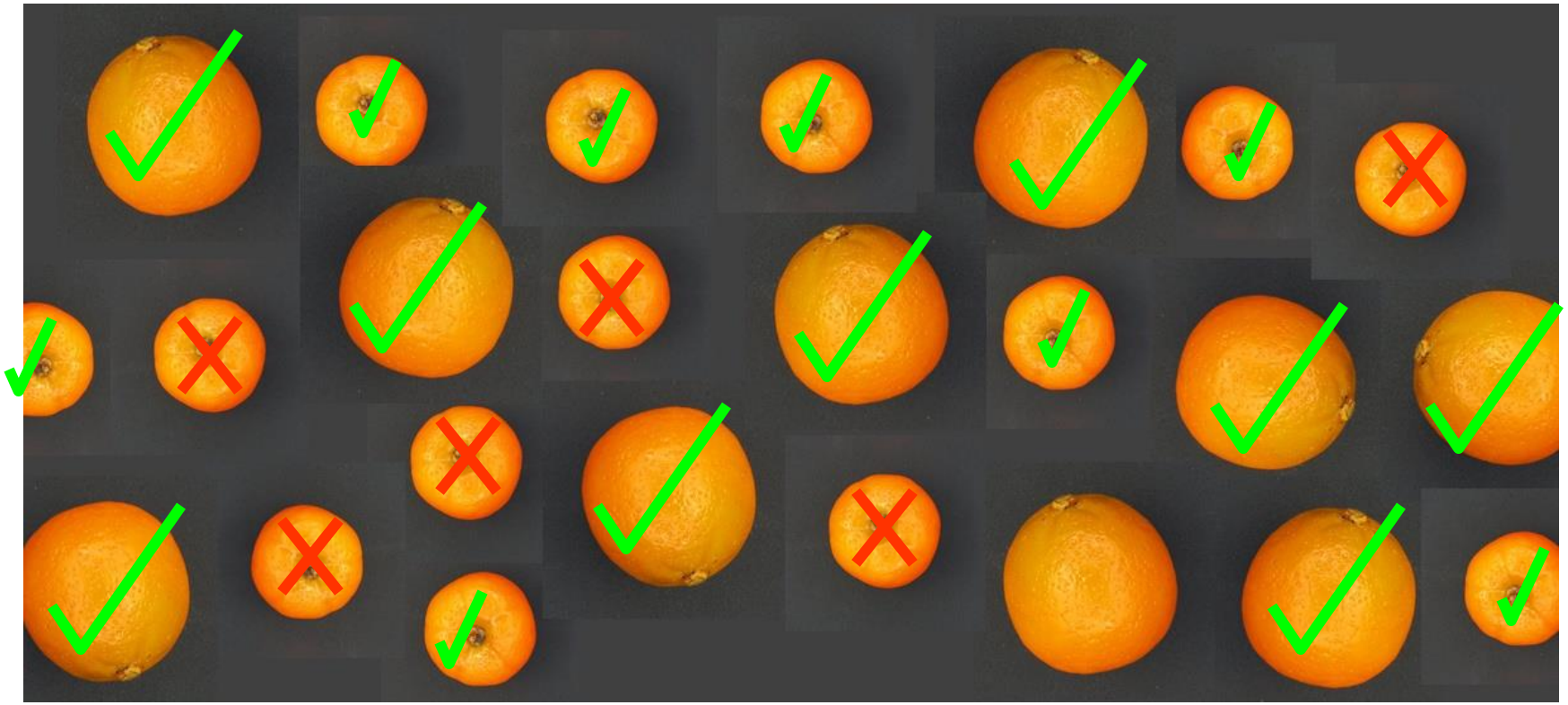
Un ejemplo práctico

Línea de decisión:



¿Y si hay muchas más naranjas que mandarinas (y valen lo mismo)?

Un ejemplo práctico



Todas las naranjas son clasificadas perfectamente
… pero el costo es que hay varias mandarinas mal clasificadas

Un ejemplo práctico

¿Cómo mejorar el desempeño?

Medición del color es una segunda alternativa: (*las naranjas son más verdes*)

Verde = 23.6%

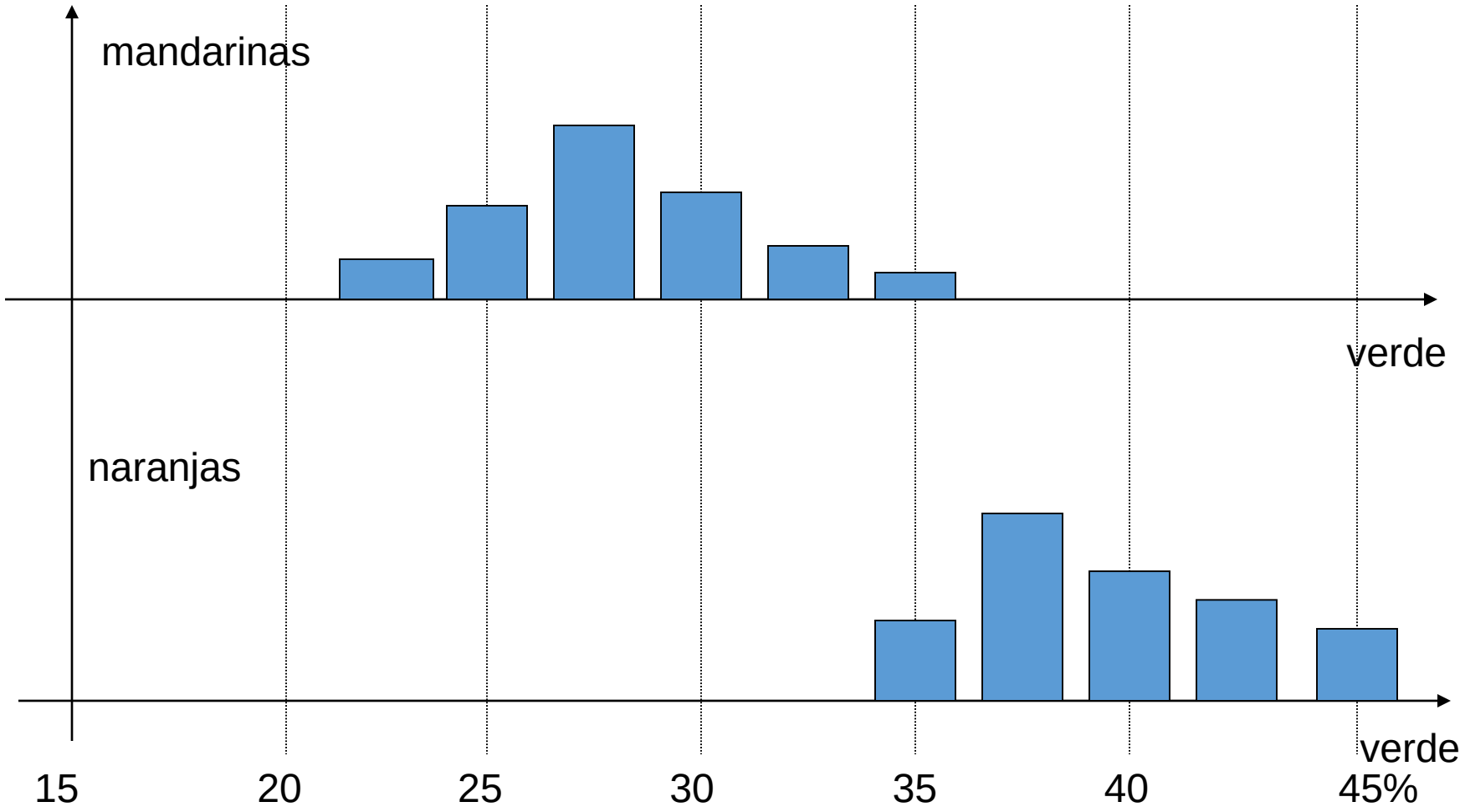
Verde = 46%

Recolección de información: Porcentaje de Verde

Naranja-01	41.3	Mandarina-01	23.6
Naranja-02	39.8	Mandarina-02	30.1
Naranja-03	36.5	Mandarina-03	37.1
Naranja-04	44.6	Mandarina-04	17.9
Naranja-05	41.2	Mandarina-05	19.7
Naranja-06	44.9	Mandarina-06	30.5
Naranja-07	44.4	Mandarina-07	35.4
:		:	
Naranja-75	38.7	Mandarina-50	33.6

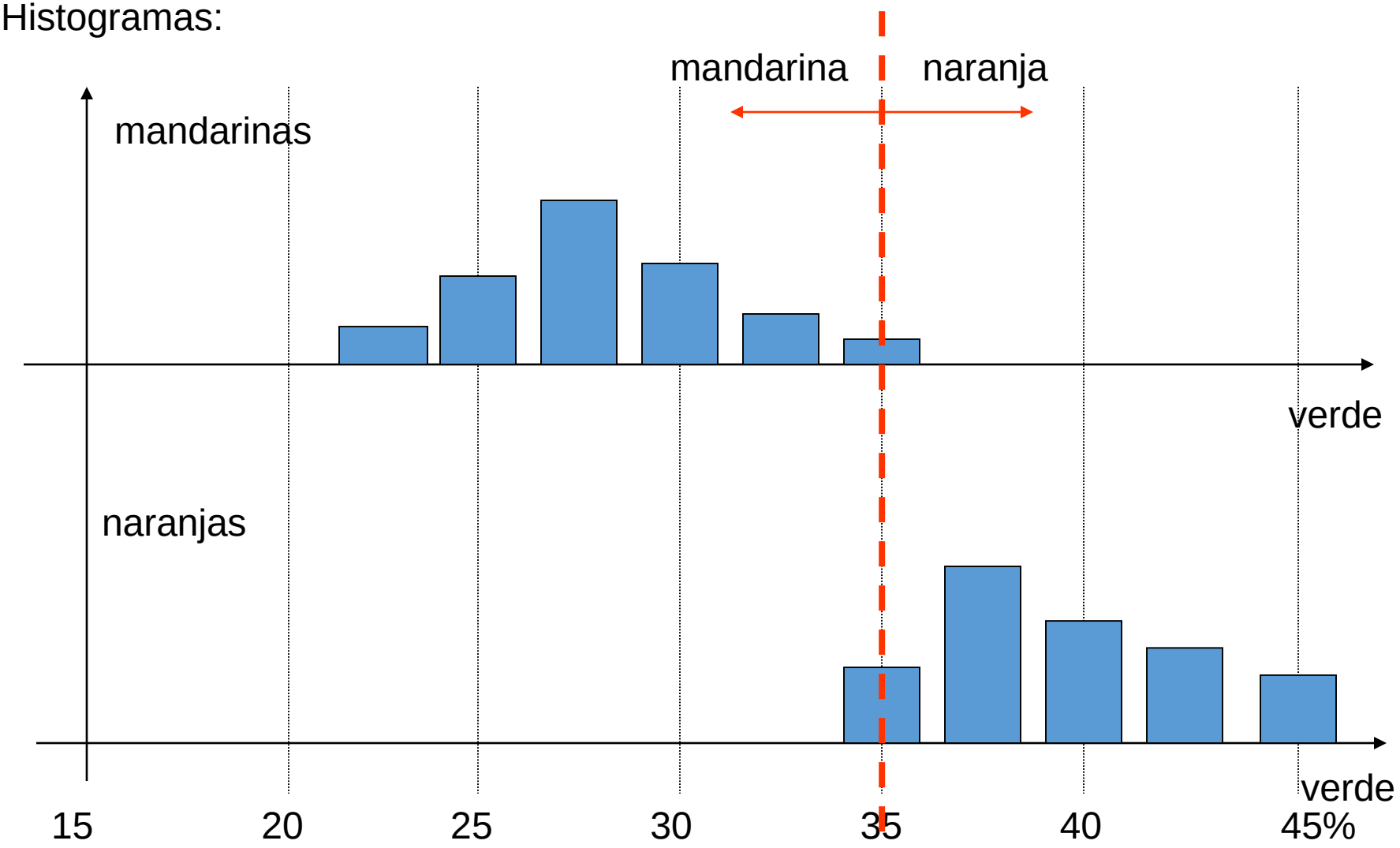
Un ejemplo práctico

Histogramas:



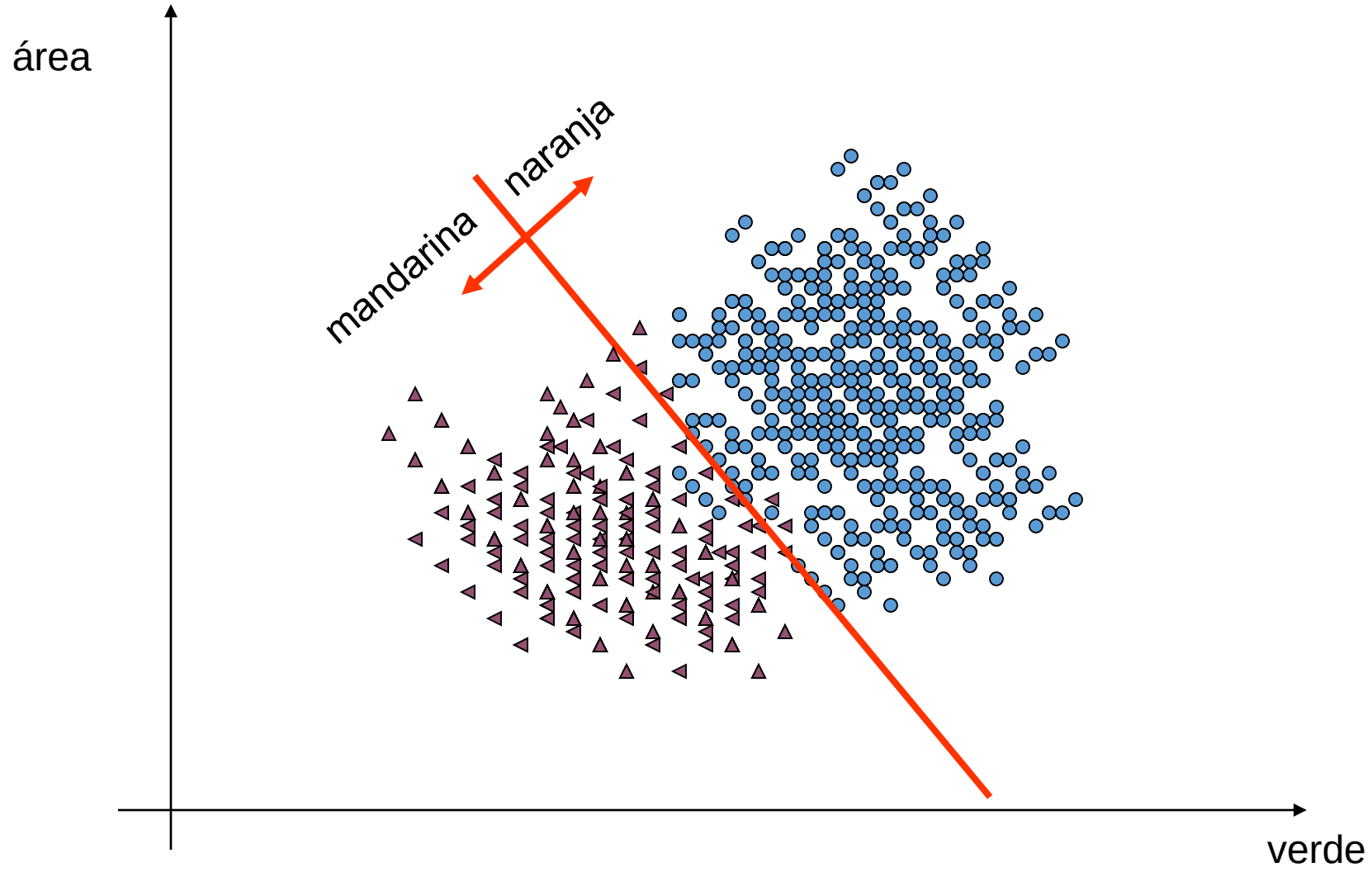
Un ejemplo práctico

Histogramas:



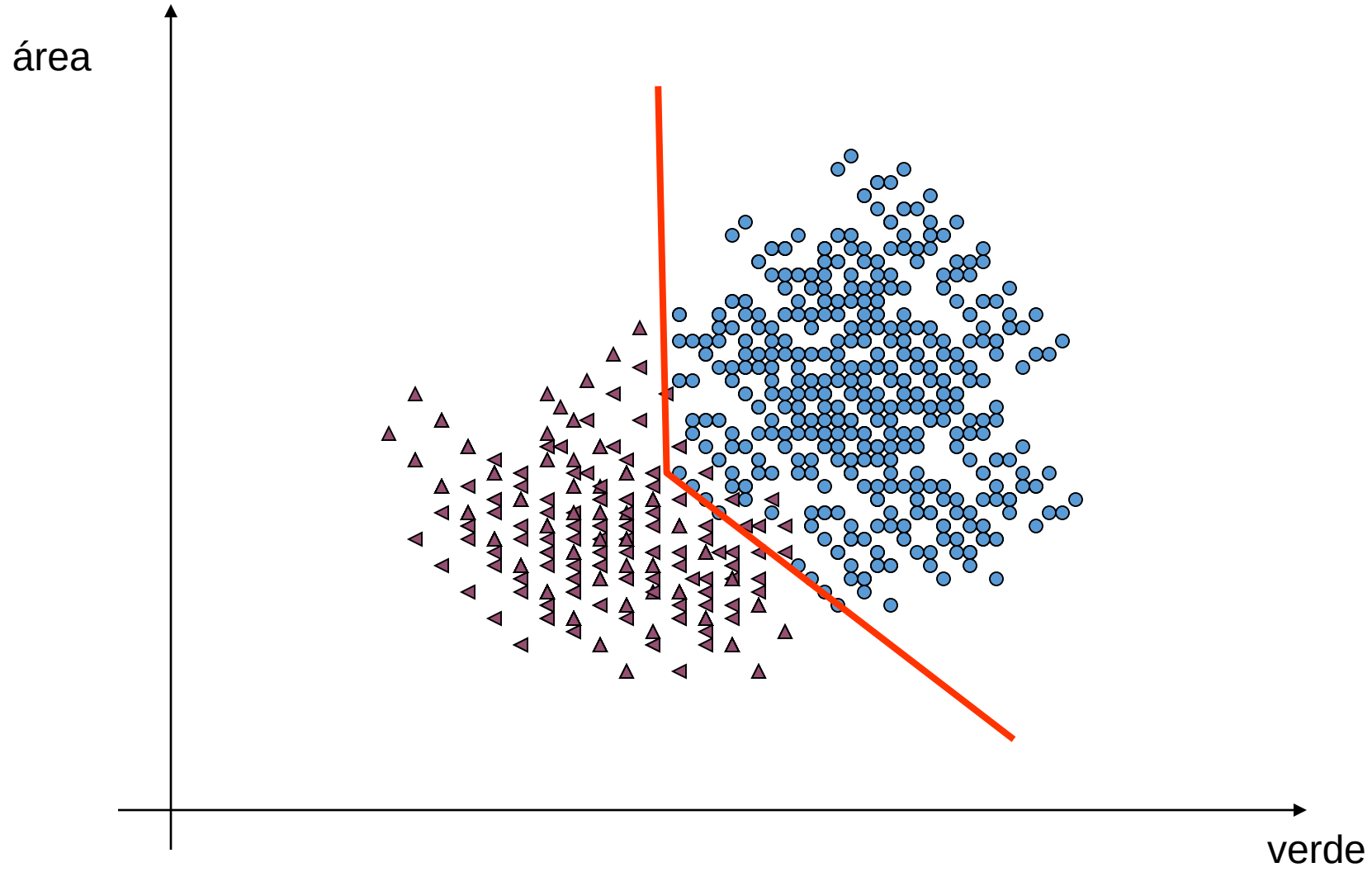
Un ejemplo práctico

Uso de dos características:



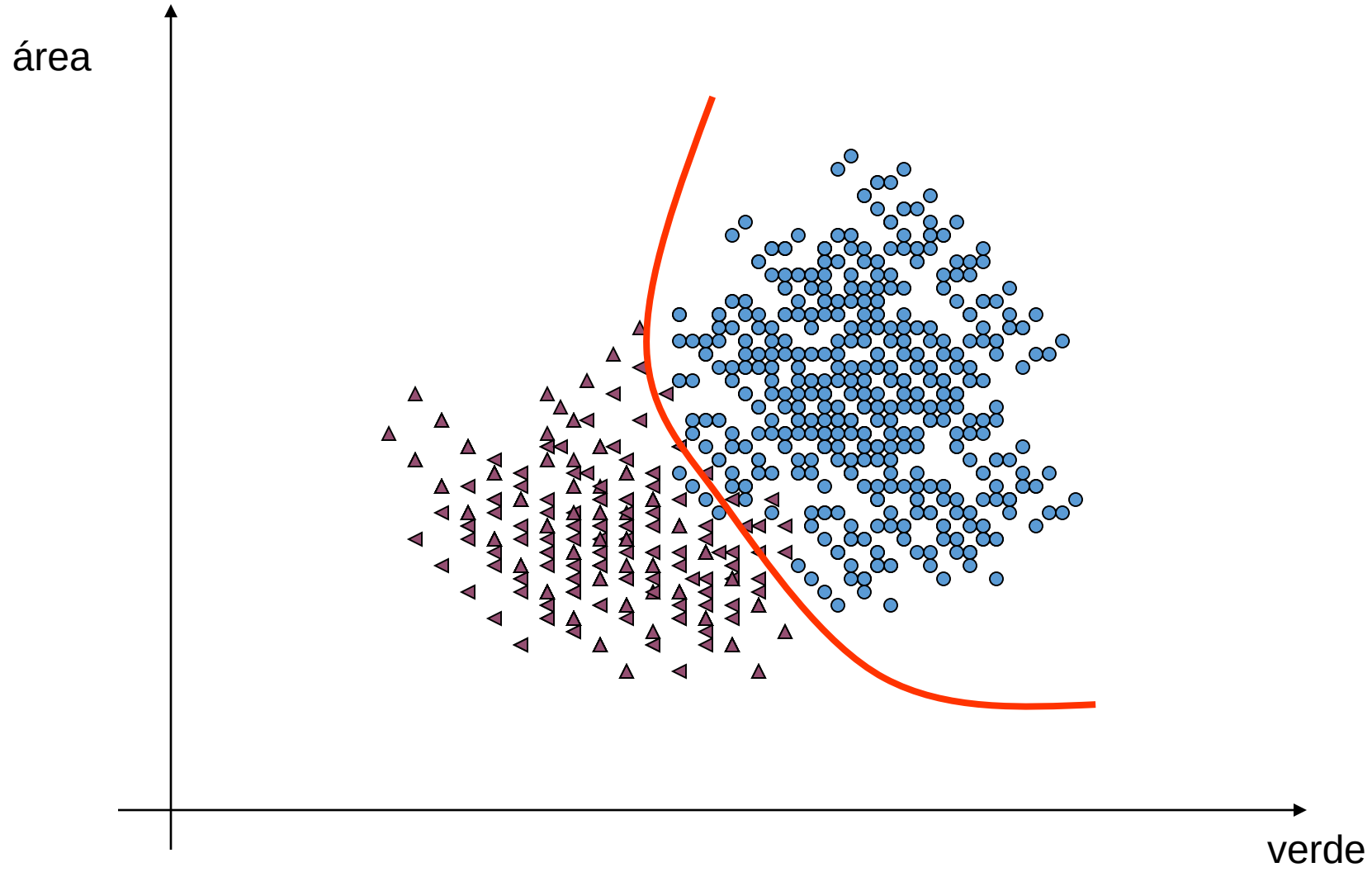
Un ejemplo práctico

Uso de dos características:



Un ejemplo práctico

Uso de dos características:



Un ejemplo práctico

Uso de dos características: