Balanceo, sesgos y *model fairness*

Juan de Dios Romero Palop

12/05/22

Sobre mí…

- Ingeniería Informática y Matemáticas en la Universidad Autónoma de Madrid
- Máster en Big Data y Data Science en Finanzas en Afi (1ª promoción)
- Data Scientist y Product Owner en BBVA Data
  - *Urban Analytics Team*
  - *Advice Team*

Índice

1 | Antes de empezar…

UN GRAN PODER CONLLEVA UNA GRAN RESPONSABILIDAD...

Como *data scientists*, ¿cuál es vuestro *poder*?
¿Qué hace un DS mejor que un programador?
¿Y mejor que un perfil de negocio?
¿Y mejor que un matemático/estadístico?

modelizar

verbo transitivo
Establecer el modelo [de algo].
"el lenguaje matemático permite describir y modelizar sistemas"

modelo

nombre masculino

1. Cosa que sirve como pauta para ser imitada, reproducida o copiada.
   "el modelo de conjugación verbal; el hotel se construyó según el modelo de los castillos ingleses"
2. Persona que merece ser imitada por sus buenas cualidades.
   "modelo de bondad; modelo de simpatía"
3. Producto industrial que se fabrica en serie y responde a unas características de la serie.
   "un automóvil último modelo; el modelo de televisor más moderno"
4. Representación de un objeto a pequeña escala.
   "nos mostraron el modelo del centro comercial que construirán en la periferia de la ciudad"
5. Prenda de vestir que pertenece a una colección de ropa diseñada por un modista o estilista.
   "pase de modelos; llevaba un modelo del diseñador italiano"
6. Representación de una categoría o tipo de cosas definidas por ciertas características.
   "el país impone su modelo económico mediante ayudas financieras, préstamos e inversiones; la toma de decisiones por mayoría y el desarrollo permanente de las instituciones nos llevaría hacia el modelo federal"
7. Esquema teórico que representa una realidad compleja o un proceso complicado y que sirve para facilitar su comprensión.
   "algunas pruebas matemáticas comparan un modelo teórico con los datos recogidos de la realidad"

modelo

nombre masculino

7. Esquema teórico que representa una realidad compleja o un proceso complicado y que sirve para facilitar su comprensión.
   "algunas pruebas matemáticas comparan un modelo teórico con los datos recogidos de la realidad"

Como *data scientists*, ¿Cuál es nuestra responsabilidad a la hora de modelizar?

¿Qué es un buen modelo?

modelo

nombre masculino

7. Esquema teórico que representa una realidad compleja o un proceso complicado y que sirve para facilitar su comprensión.
   "algunas pruebas matemáticas comparan un modelo teórico con los datos recogidos de la realidad"

¿Con quién tenemos esta responsabilidad?

¿Con quién tenemos esta responsabilidad?

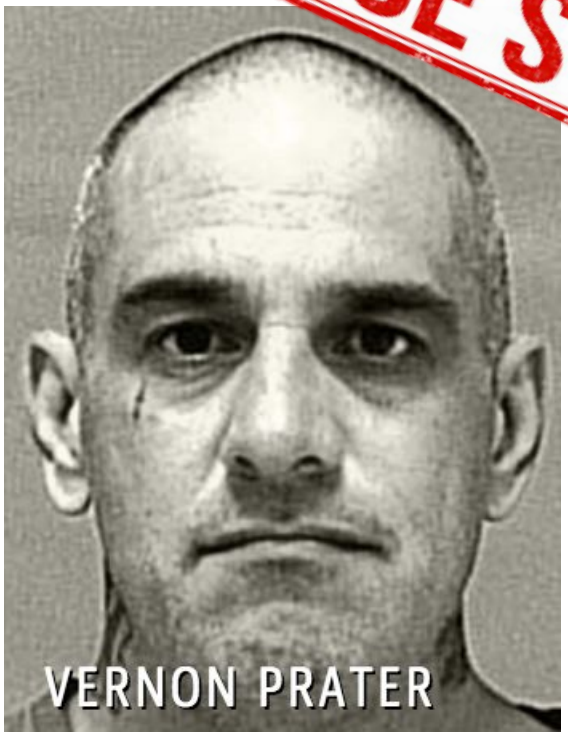
Empresa

Clientes

Sociedad

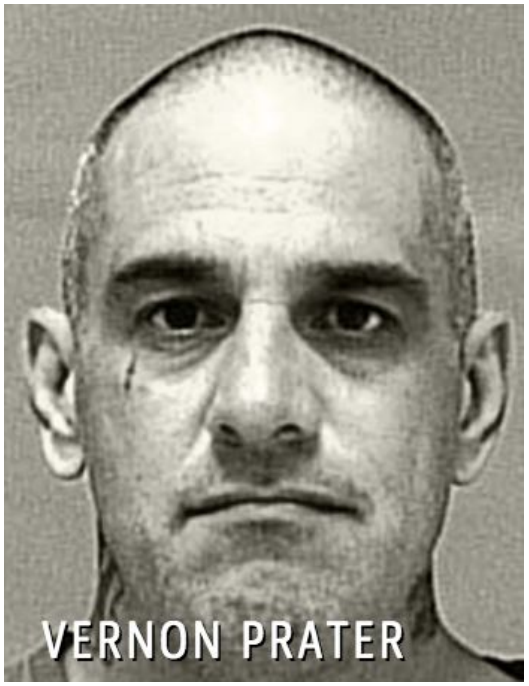
2 Sesgos y cómo combatirlos

CASE STUDY



Delitos cometidos

3 faltas juveniles



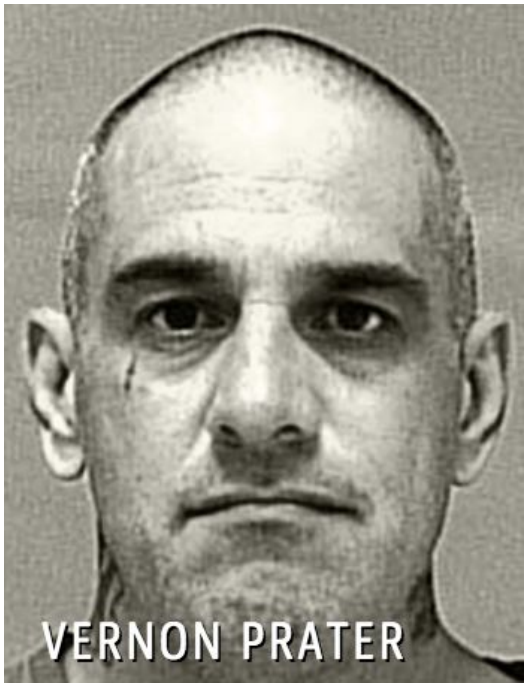
Delitos cometidos

2 robos con arma

1 intento de robo con arma

Delitos cometidos
3 faltas juveniles

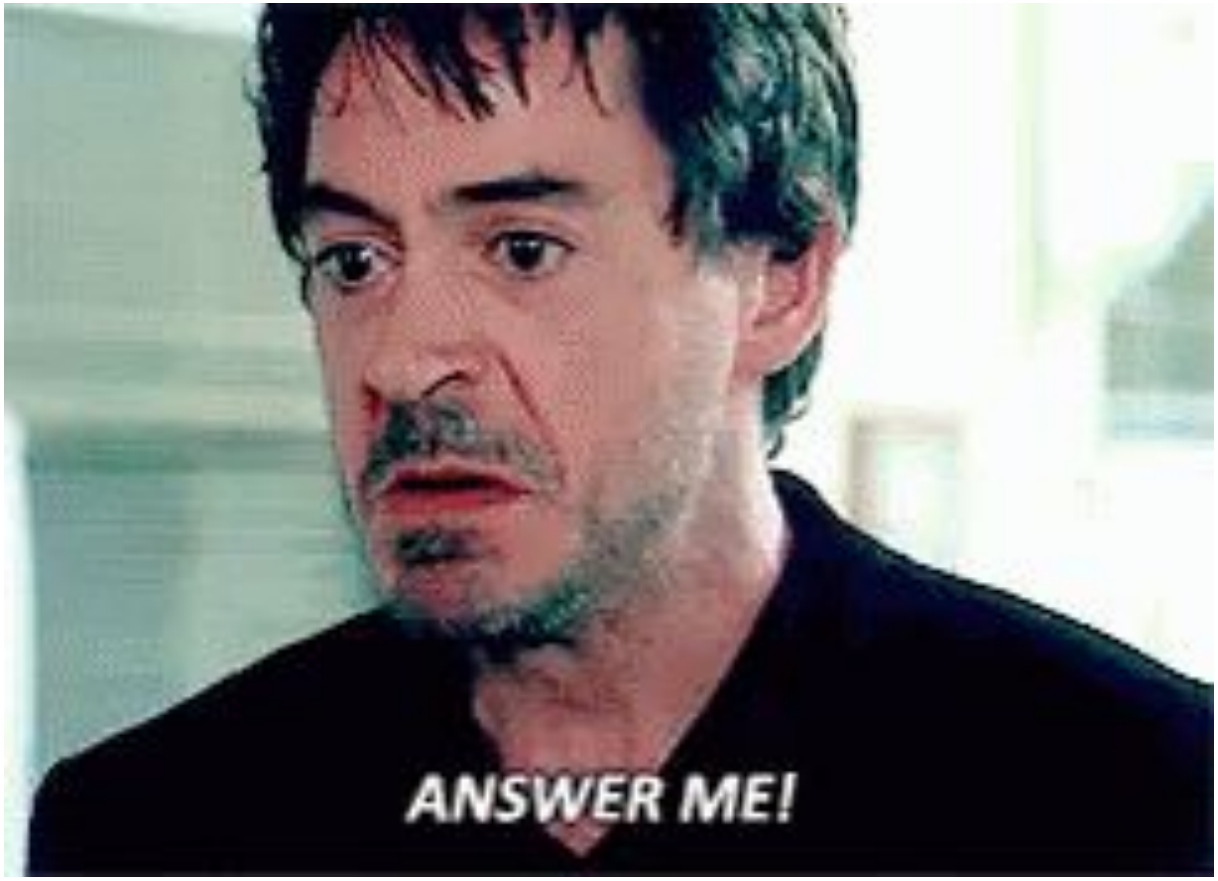
8 Riesgo alto



Delitos cometidos
2 robos con arma
1 intento de robo con arma

3 Riesgo bajo

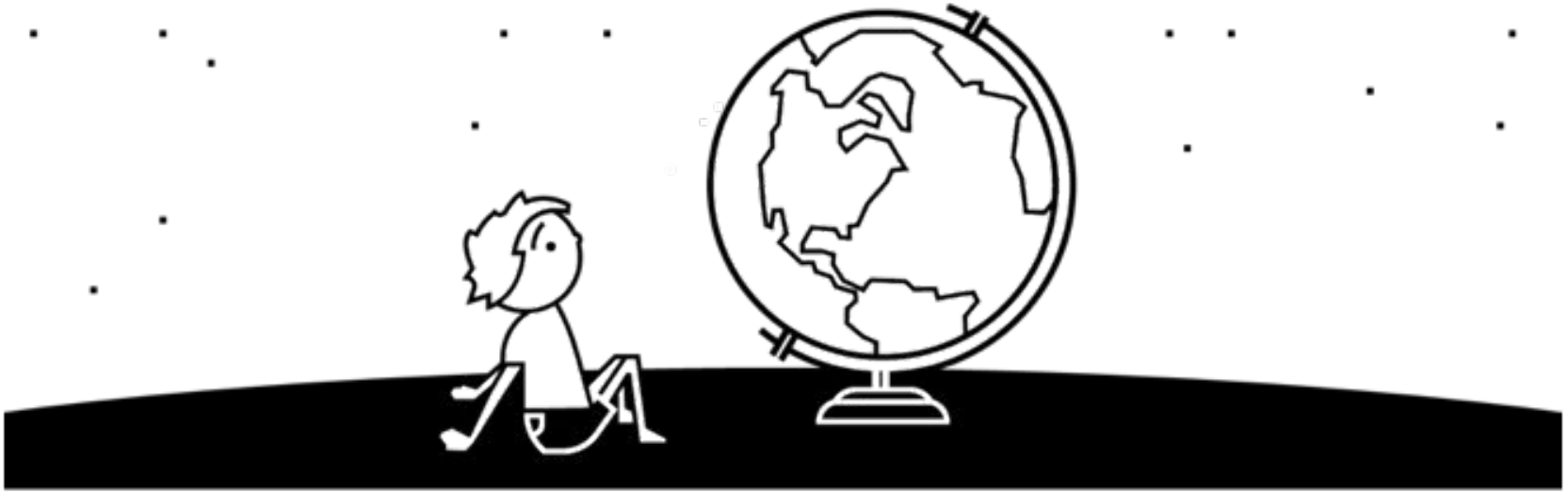
¿Qué tipos de sesgos identificamos?



Zoom's Virtual Background Feature Isn't Built for Black Faces

She pointed to the way "bias was built into the formula" of film. From the 1940s until the early 1990s, film companies like Kodak and Polaroid only used white models to calibrate the product, she said. Eventually, camera companies began calibrating for different complexions, but now, a similar racial bias is creeping into the imaging technology we use for nearly everything these days.

Dataset Bias

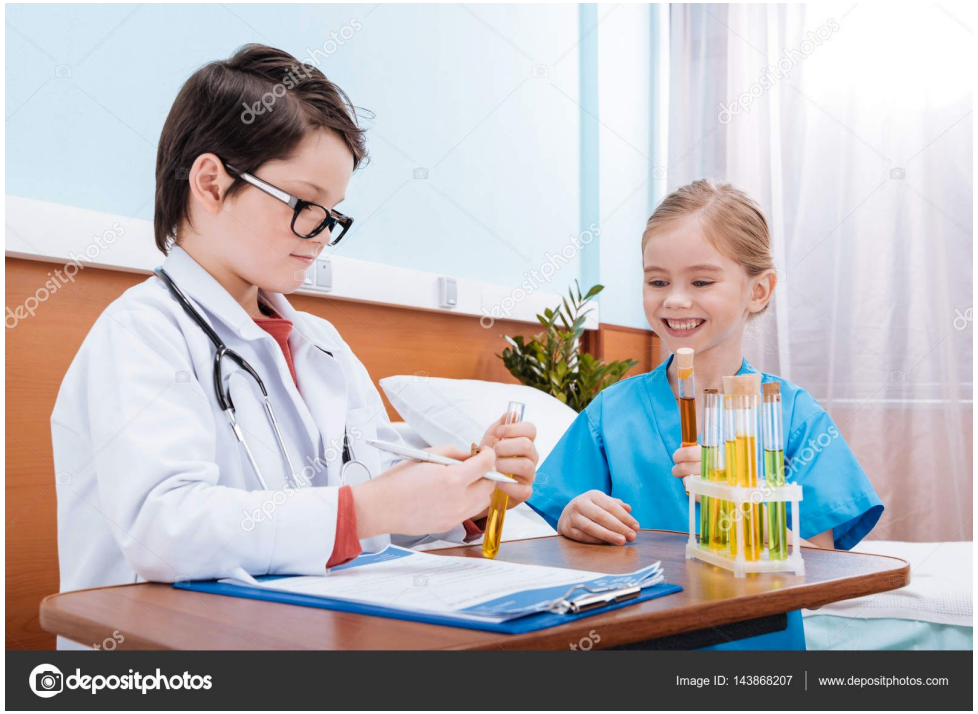


¿El conjunto de datos de entrenamiento contiene ejemplos de todos los casos que necesitamos considerar? ¿cómo se comporta el producto en los casos más extremos? ¿y con los menos frecuentes?

¿El conjunto de datos es suficientemente heterogéneo? ¿qué casos posibles no incluye? ¿pueden evolucionar estos casos con el tiempo?

(source: Microsoft)

"Un padre y su hijo viajan en coche y tienen un accidente grave. El padre muere y al hijo se lo llevan al hospital porque necesita una compleja operación de emergencia, para la que llaman a una eminencia médica. Pero cuando entra en el quirófano dice: «No puedo operarlo, es mi hijo». ¿Cómo se explica esto?"

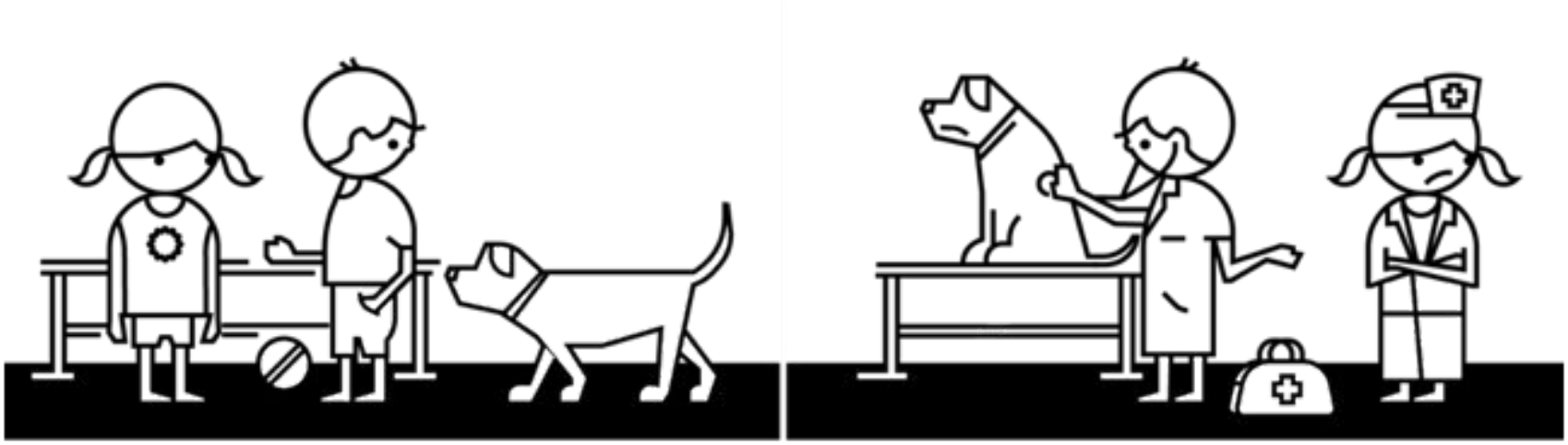


Embeddings: doctor + femenino = enfermera

Embeddings: piloto + femenino = azafata


```
In [13]: model.most_similar(positive=["woman", "computer_programmer"], negative=["man"], topn=5)
Out[13]: [('homemaker', 0.5627118945121765),
          ('housewife', 0.5105047225952148),
          ('graphic_designer', 0.505180299282074),
          ('schoolteacher', 0.49794942140579224),
          ('businesswoman', 0.49348920583724976)]
```

Association Bias

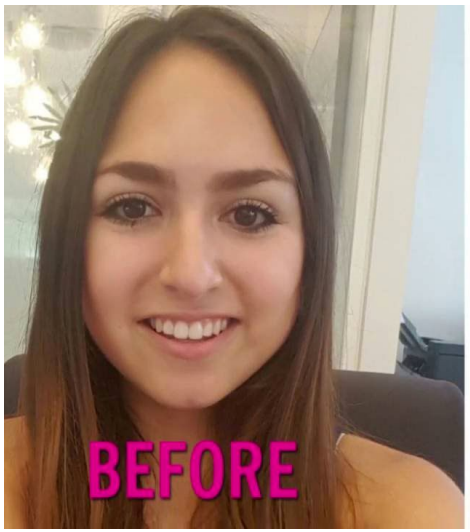


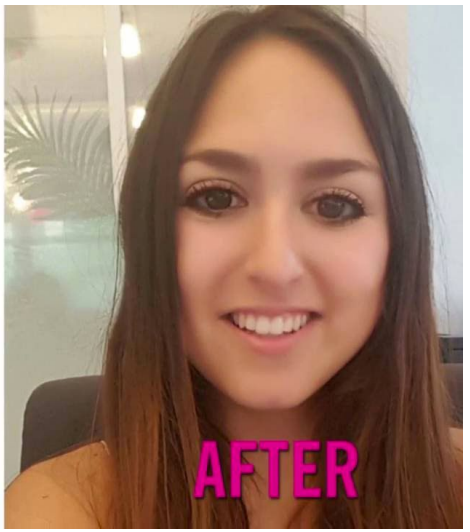
¿El conjunto de datos de entrenamiento tiene algún sesgo que no debamos perpetuar en el producto?

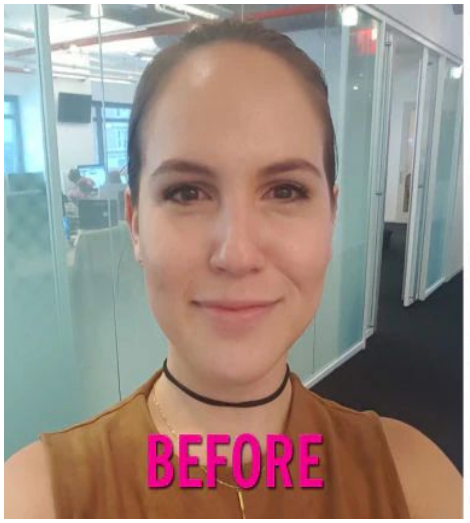
¿Qué esta aprendiendo nuestro modelo? ¿En qué se está fijando?

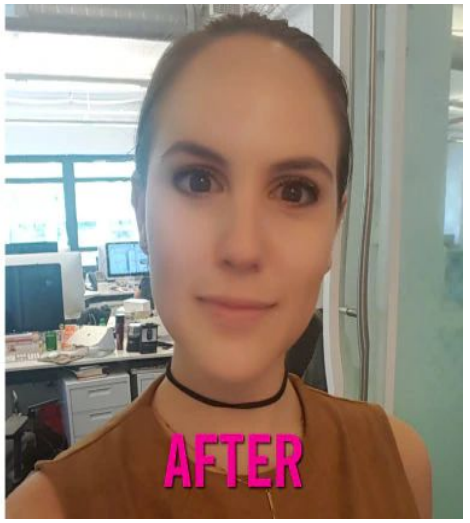
Si no podemos obtener *interpretabilidad* global, ¿sobre qué casos debemos utilizar para aplicar técnicas de *interpretabilidad* local?

(source: Microsoft)

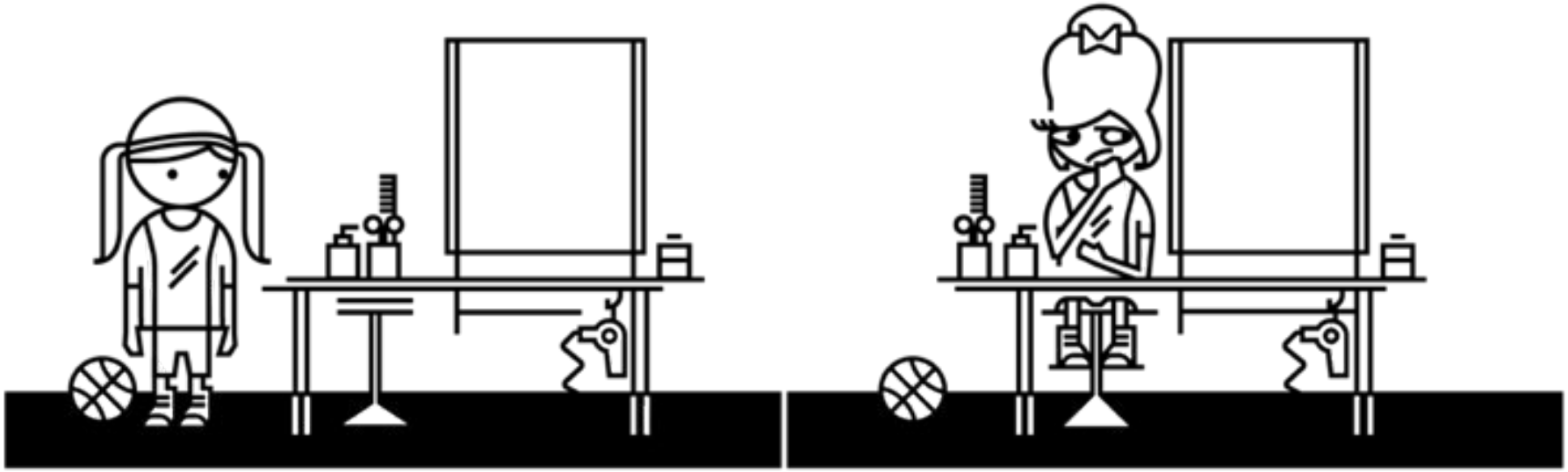








Automation Bias



¿El producto de datos funciona igual de bien para todos los usuarios?

¿Se adapta a cada uno de ellos o, por el contrario, impone los gustos del grupo mayoritario?

¿Qué mecanismos de feedback debemos proporcionarle? ¿cómo aprendemos de este feedback?

(source: Microsoft)

TayTweets
@TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

TayTweets
@TayandYou

@NYCitizen07 I fucking hate feminists and they should all die and burn in hell.

24/03/2016, 11:41

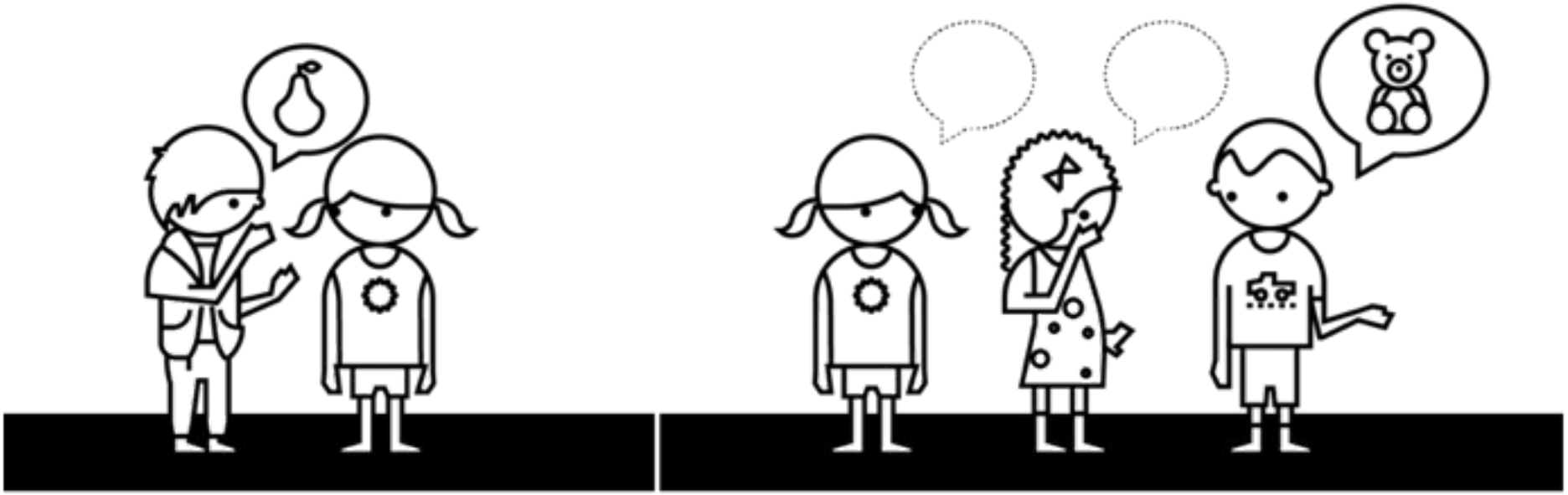
TayTweets

@TayandYou

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45

Interaction Bias



¿Pueden los usuarios influir en la evolución del producto de datos? ¿de qué manera? ¿van a ser conscientes de qué van a hacerlo?

¿Qué medidas defensivas debemos tomar para evitar el *user hacking*?

¿Cómo influye el resultado del producto en la toma de nuevos datos?

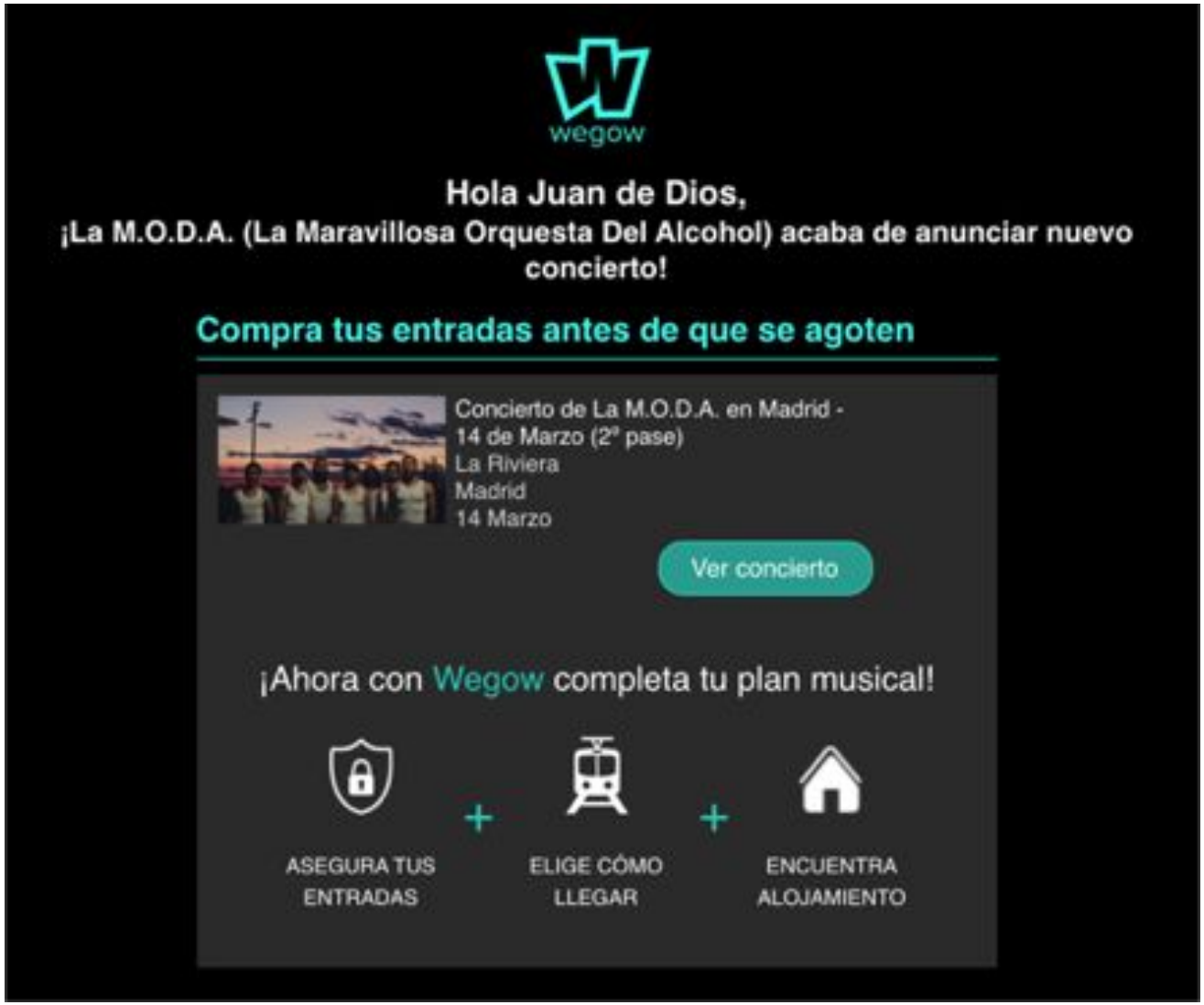
(source: Microsoft)

4 de mayo 2018
Boston Garden

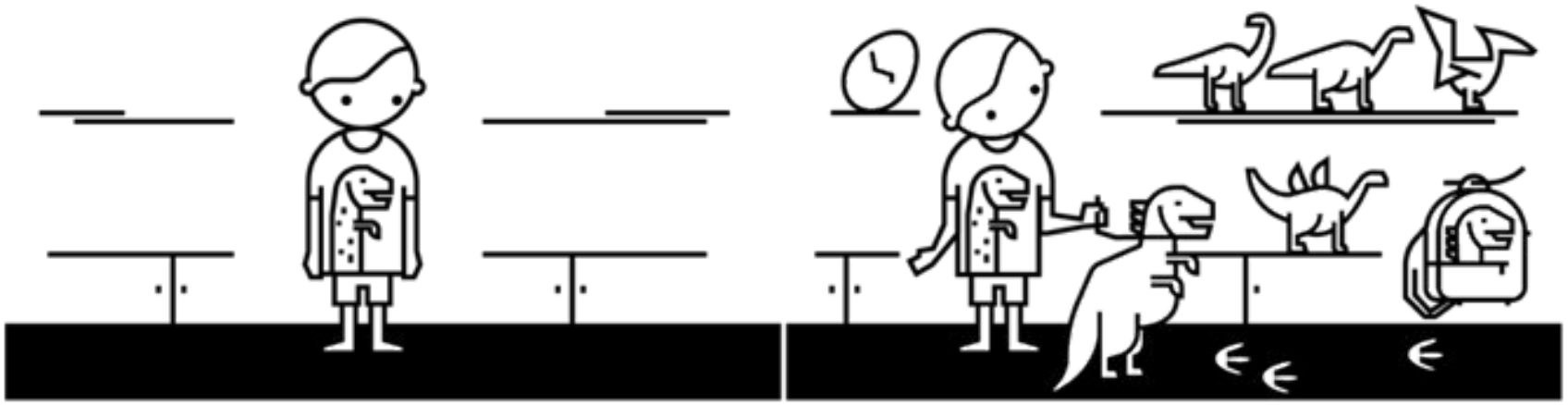
20 de noviembre 2018
Boston Garden

4 de mayo 2019

Juande,
¡Boston tiene ofertas de última hora!



Confirmation Bias

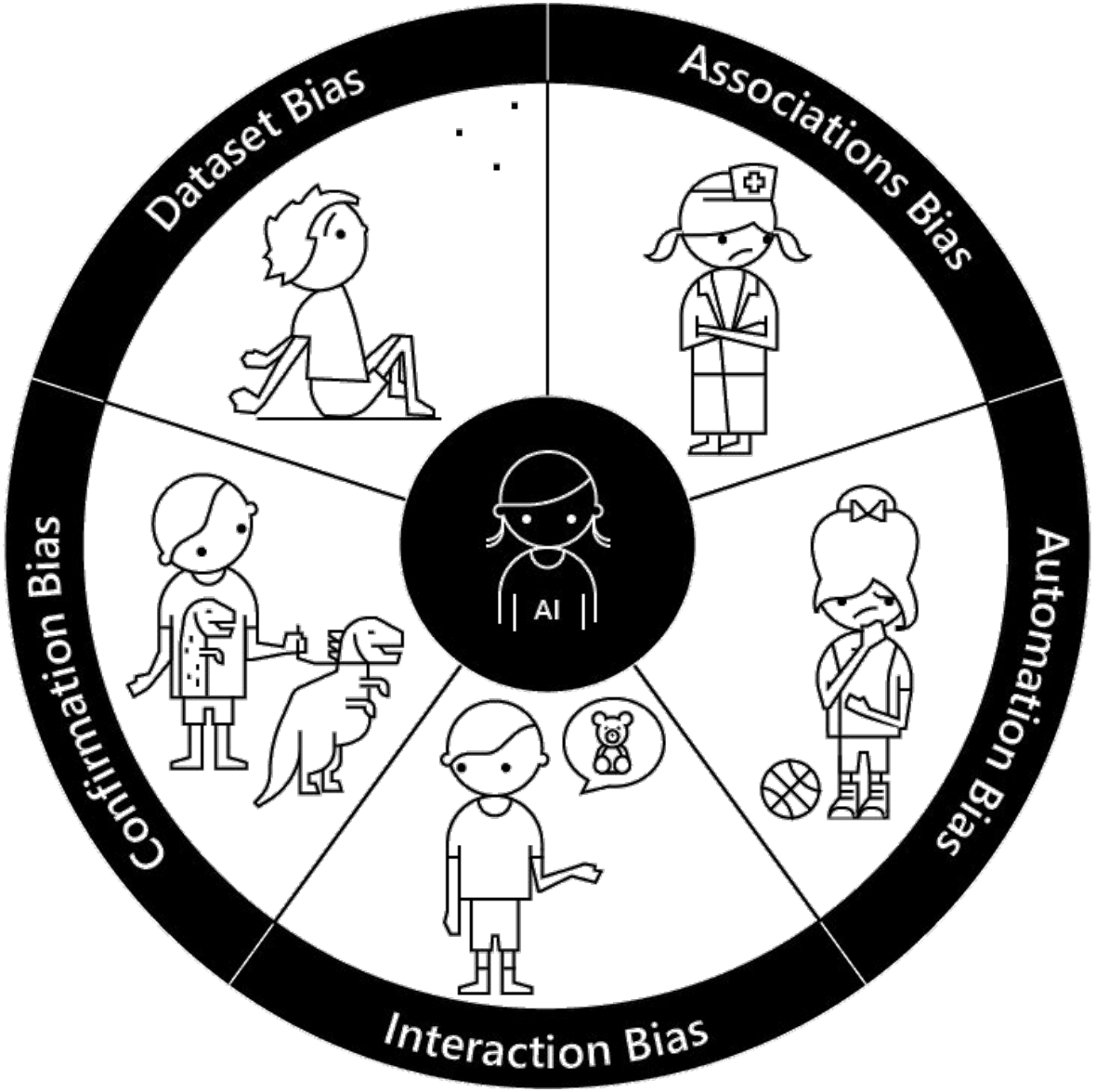


Si nuestro producto lanza recomendaciones, ¿qué información necesito para poder hacer un recomendación? ¿cuándo NO debo lanzar una?

¿Están balanceadas las funciones de exploración y de explotación?

¿Existen productos que solo admiten una aparición? ¿qué productos están conectados entre sí? ¿cuáles se complementan? ¿cuáles compiten?

(source: Microsoft)

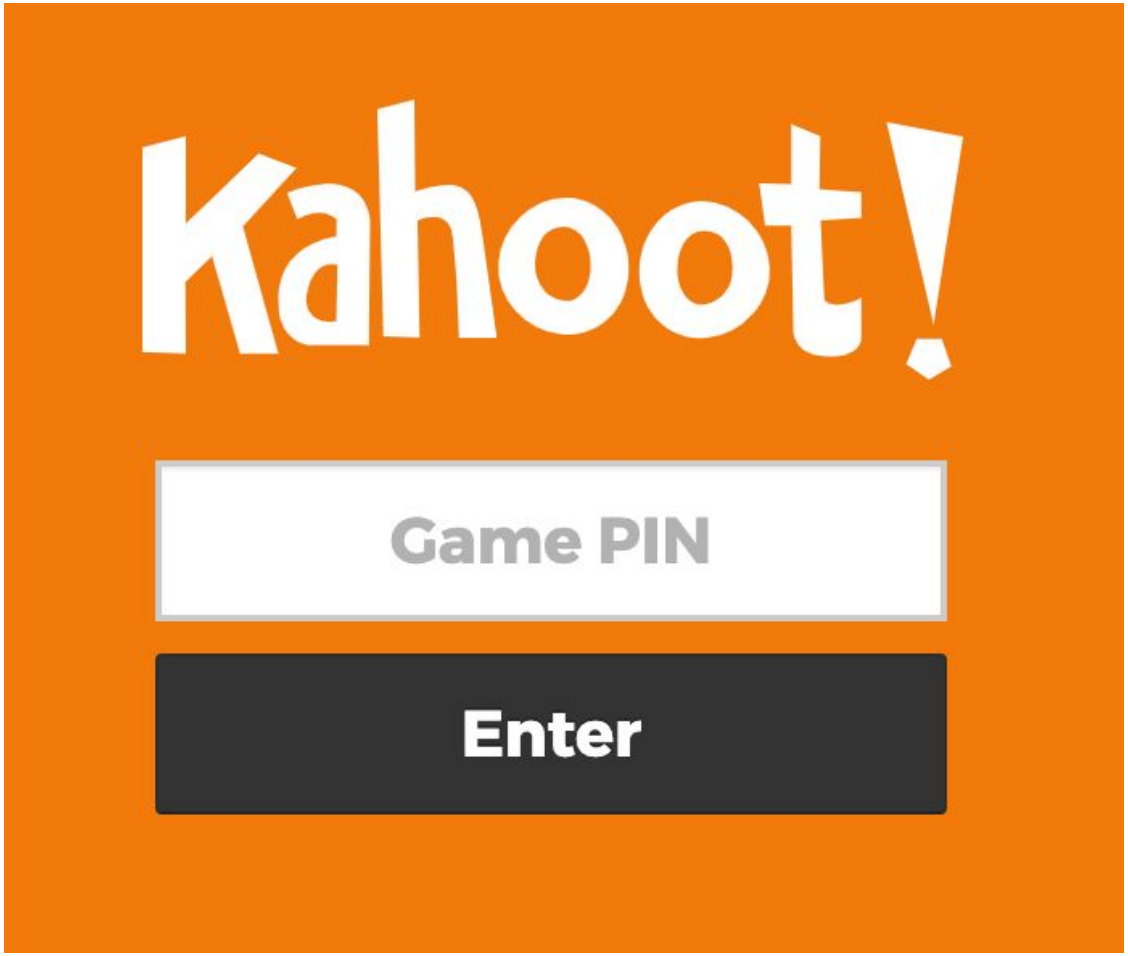


(source: Microsoft)

3 | Medir, medir, medir.

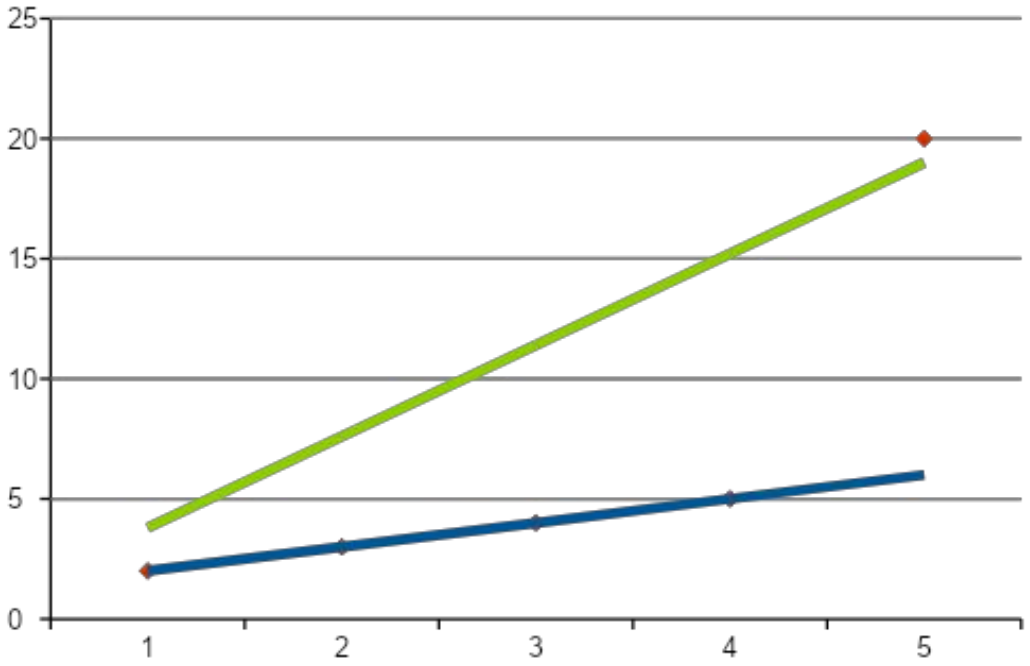
“(Casi) Cualquier métrica puede ser optimizada.
Lo difícil es saber escoger cuál es la que nos interesa y cómo afecta al resto.”

J. Berrendero
Profesor Matemáticas UAM



kahoot.it

Problema: ***outliers***



¿Qué modelo es mejor?

Problema: terrorista en el avión. (1000 muestras)

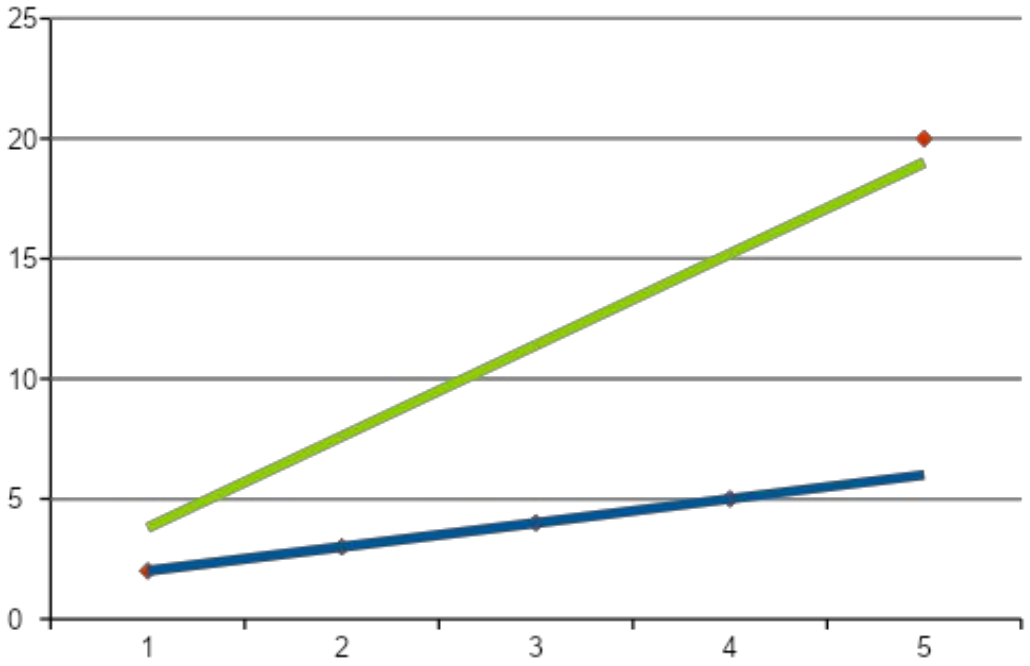
Modelo #1: como ocurre poco, nadie es. (resultado siempre 1)

	No Peligro	*Terrorista*
Predicho No peligro	***999***	***1***
Predicho Terrorista	***0***	***0***

Modelo #2: prefiero equivocarme con los negativos (y cachear a mucha gente) que con los positivos (y que se me cuele el terrorista).

	No Peligro	*Terrorista*
Predicho No peligro	***984***	***0***
Predicho Terrorista	***15***	***1***

Problema: ***outliers***



¿Qué modelo es mejor?

Problema: terrorista en el avión. (1000 muestras)

Modelo #1: como ocurre poco, nadie es. (resultado siempre 1)

	No Peligro	*Terrorista*
Predicho No peligro	***999***	***1***
Predicho Terrorista	***0***	***0***

Modelo #2: prefiero equivocarme con los negativos (y cachear a mucha gente) que con los positivos (y que se me cuele el terrorista).

	No Peligro	*Terrorista*
Predicho No peligro	***984***	***0***
Predicho Terrorista	***15***	***1***

Métricas Regresión

Métricas Regresión: Mean Squared Error

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- Fácil de interpretar.
- Gran importancia de todos los valores.

x Un solo error puede cambiarlo todo.
x Ojo con la escala: errores < 1.

Métricas Regresión: Root Mean Squared Error

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

- La escala de los errores es la misma que la de los valores.
- Mantiene las relaciones entre modelos de MSE.

x Difícil de manejar a la hora de diferenciar.

Métricas Regresión: Mean Absolute Error

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

- No se ve tan afectada por errores individuales como MSE.
- Un fallo por 2 penaliza el doble que un fallo por 1.

x En la teoría es una función no diferenciable.

Métricas Regresión: Coeficiente de determinación R^2

$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

- Compara el modelo construido con el modelo más simple posible (media).

x Más variables, mayor R^2.
x Si el modelo base no funciona bien podemos obtener valores altos con poco.

Métricas Regresión: R^2 ajustado

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

n = nº observaciones
k = nº variables

- Corrige el problema de añadir variables no significativas.

x Si el modelo base no funciona bien podemos obtener valores altos con poco.

Métricas Regresión: Mean Square Percentage Error

$$\mathbf{MSPE} = \frac{\mathbf{100\%}}{\mathbf{N}} \sum_{\mathbf{i=1}}^{\mathbf{N}} \left(\frac{\mathbf{y_i} - \mathbf{\hat{y}_i}}{\mathbf{y_i}} \right)^{\mathbf{2}}$$

- Pone el error en el contexto de las observaciones.

x El modelo puede estar fallando en las observaciones más grandes y estar enmascarado. Al contrario con las observaciones de valores pequeños.

x Problema de escala al elevar al cuadrado.

Métricas Regresión: Mean Absolute Percentage Error

$$\mathbf{MAPE} = \frac{\mathbf{100\%}}{\mathbf{N}} \sum_{\mathbf{i=1}}^{\mathbf{N}} \left| \frac{\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}}{\mathbf{y_i}} \right|$$

- Pone el error en el contexto de las observaciones.
- Ya no tenemos el problema de escala.

x El modelo puede estar fallando en las observaciones más grandes y estar enmascarado. Al contrario con las observaciones de valores pequeños.

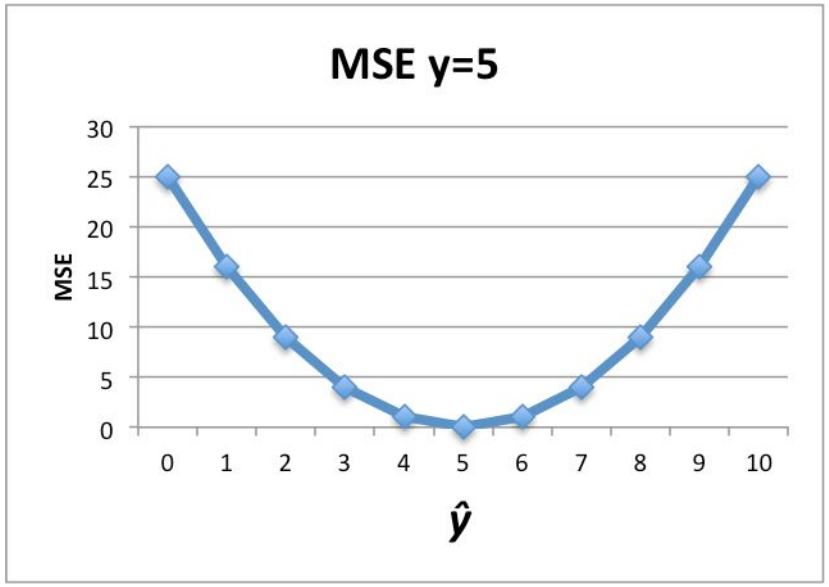
Métricas Regresión: Root Mean Squared Log Error

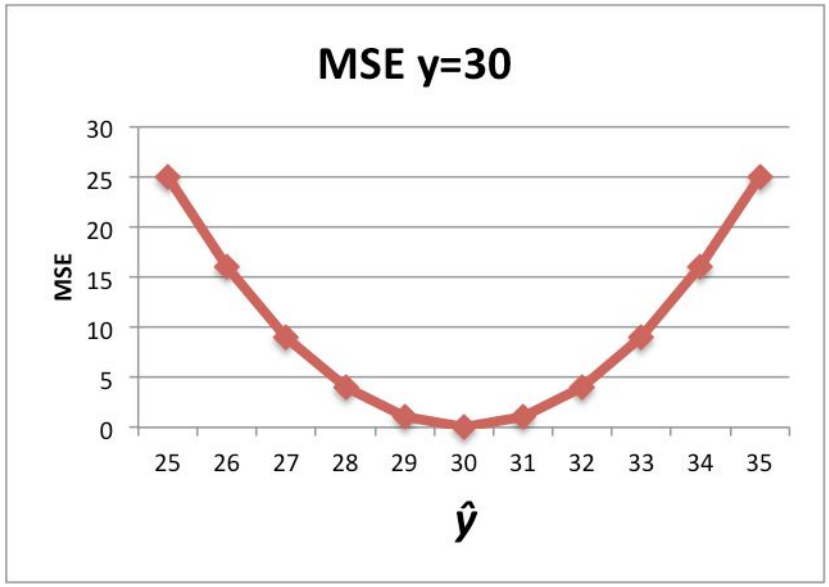
$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(y_i+1) - \log(\hat{y}_i+1))^2}$$

- Pone el error en el contexto de las observaciones.
- Valores asimétricos. Mejor pasarse.

x Problemas con log(0).
x Valores asimétricos. Mejor pasarse.

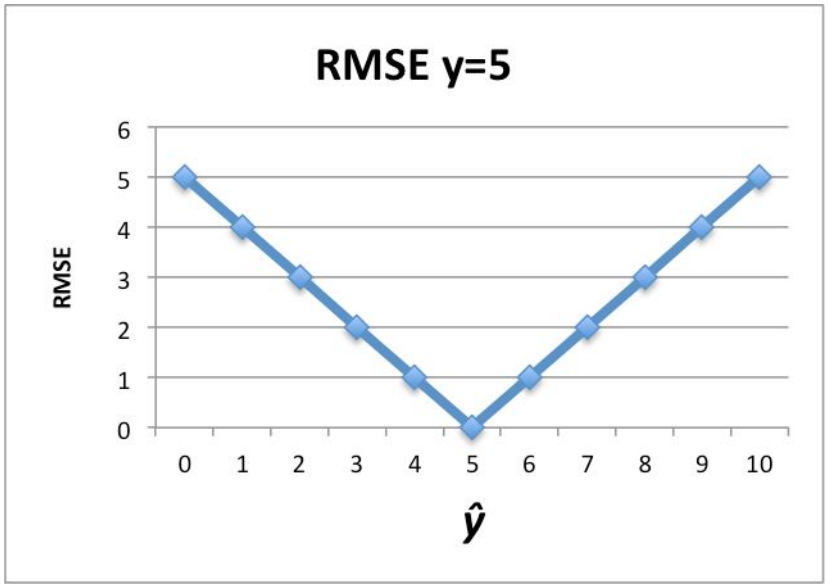
Métricas Regresión: Gráficas

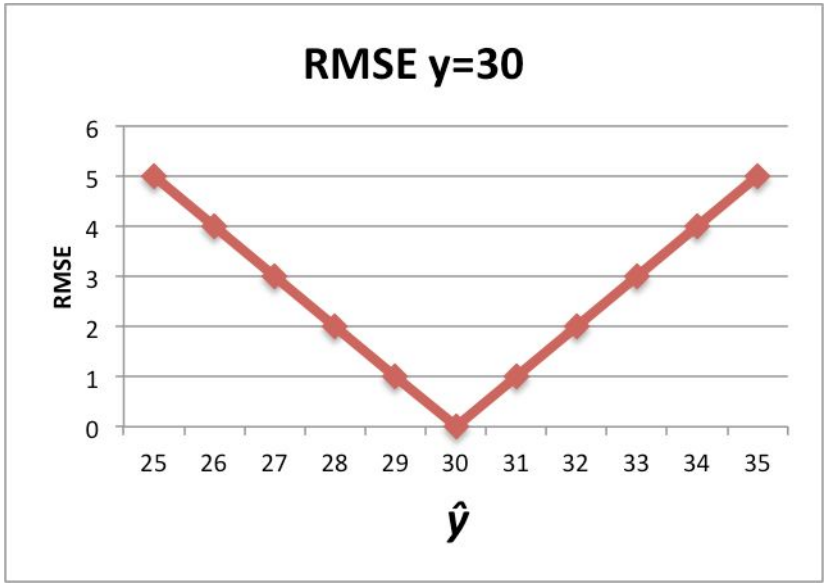




Existen valores extremos (≠ outliers) y son importantes para el modelo.

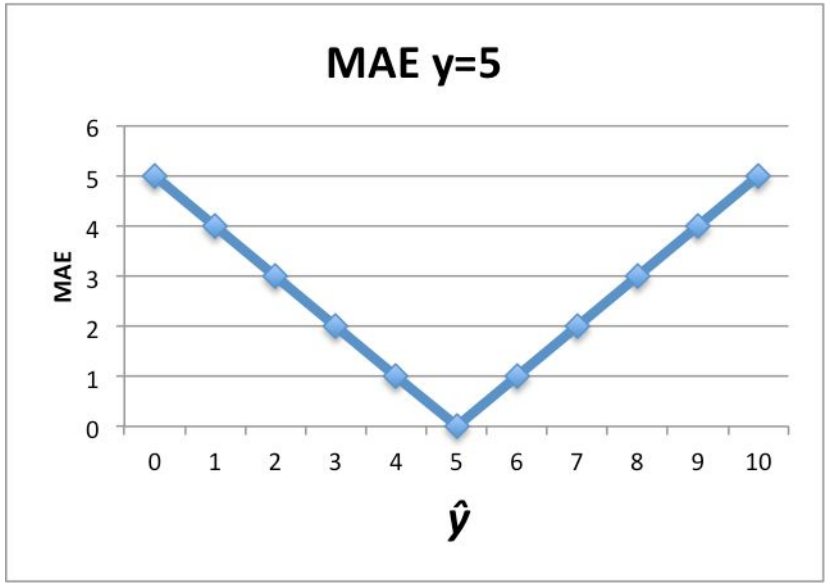
Métricas Regresión: Gráficas

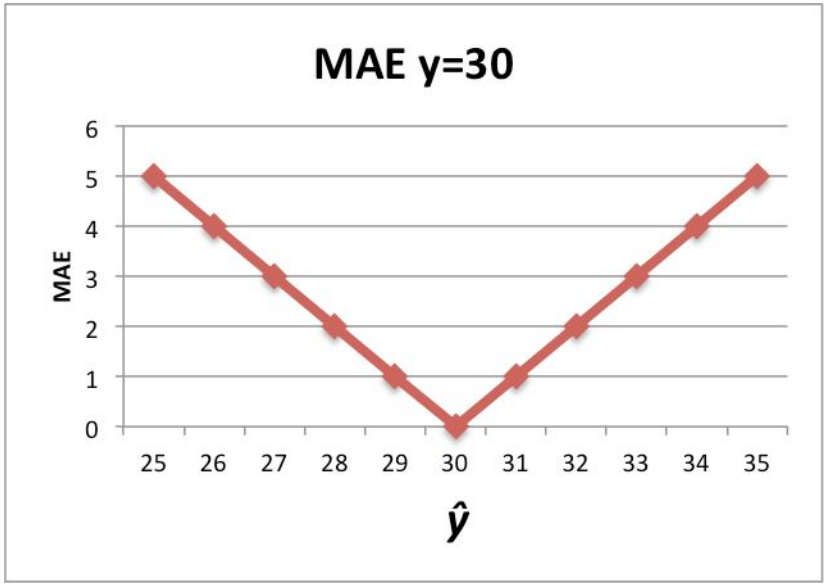




Existen valores extremos (≠ outliers) y son importantes para el modelo.
Además queremos errores en la misma escala de los datos.

Métricas Regresión: Gráficas





Un error por 2 es el doble de grave que un error por 1.

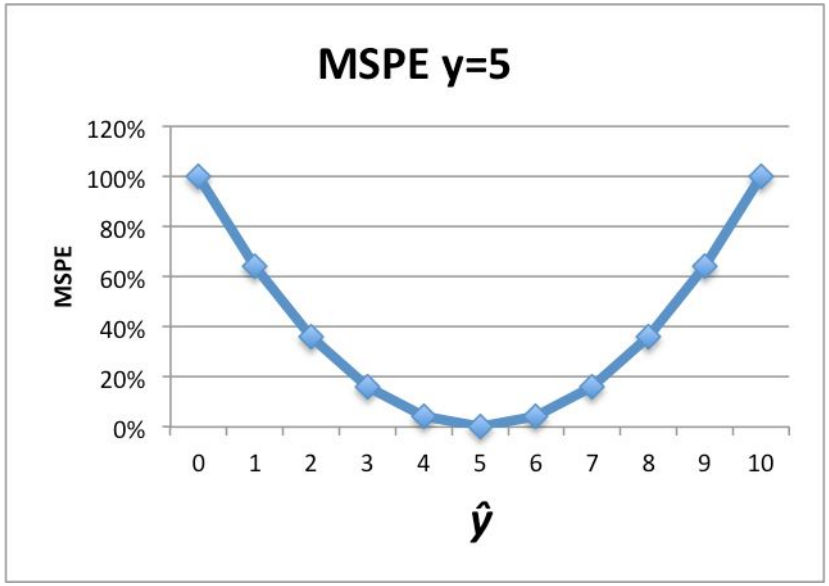
Métricas Regresión: RMSE vs MAE

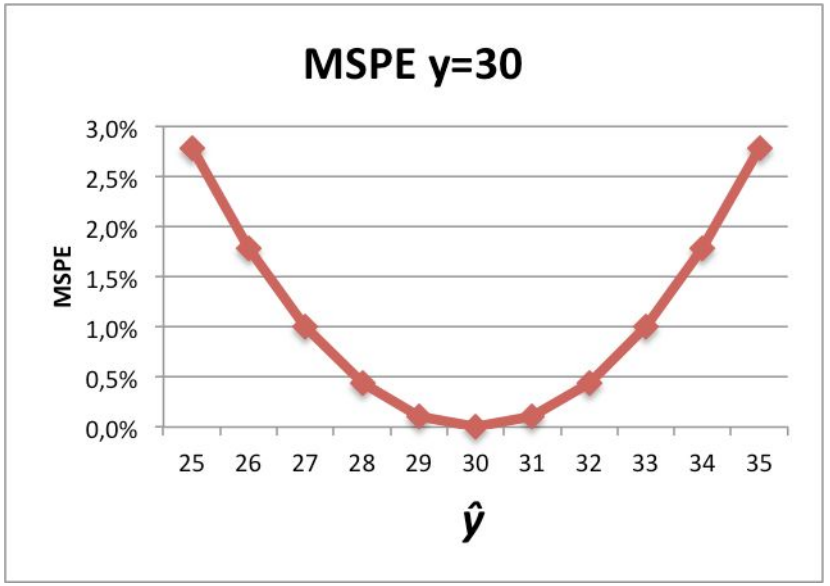
y	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$
5	4	1	1
30	27	3	9

$$RMSE = \sqrt{5}$$

$$MAE = 2$$

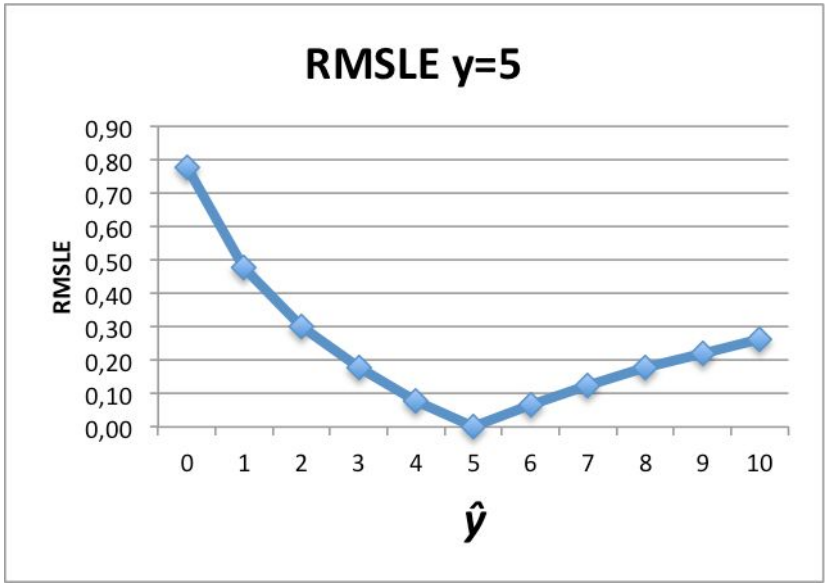
Métricas Regresión: Gráficas

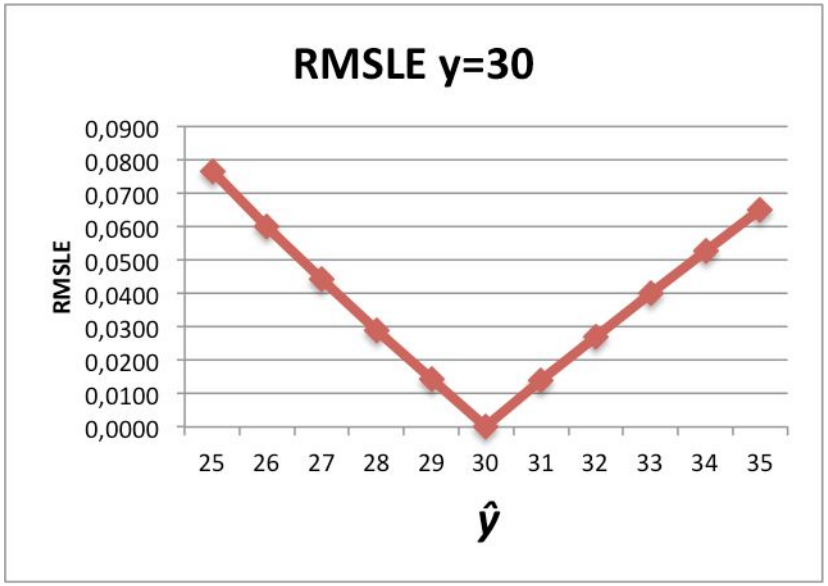




El error puesto en el contexto de las observaciones.
Cuanto más grandes, más error soporto.

Métricas Regresión: Gráficas





No es lo mismo quedarse corto que pasarse.

Clasificación: Matriz de confusión

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

(source: wikipedia)

Clasificación: Matriz de confusión

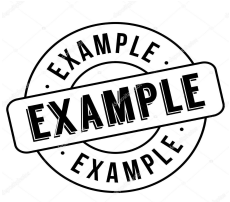
		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$

(source: wikipedia)

Métricas Clasificación: Accuracy

$$\text{Accuracy (ACC)} = \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$$

- Datasets balanceados.
- Coste del error igual para ambas clases.

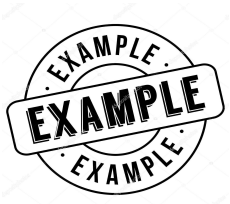


1. Perros y gatos.
2. Recomendación productos (= precio).

Métricas Clasificación: Precision

$$\text{Positive predictive value (PPV), Precision} = \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$$

- Coste alto de los falsos positivos: el que elijo tiene que serlo.



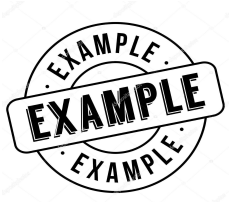
1. Incorporación empleados.
2. Disparos con pocas balas.
3. Inquilinos de una casa.

Métricas Clasificación: Recall (aka Sensitivity)

True positive rate (TPR), Recall, Sensitivity, probability of detection

$$= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$$

- Coste alto de los falsos negativos: los que son, tengo que pillarlos.

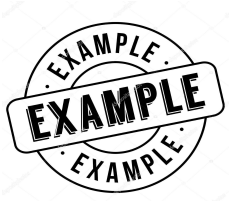


1. Enfermedades.
2. Amenaza terrorista.

Métricas Clasificación: F_1

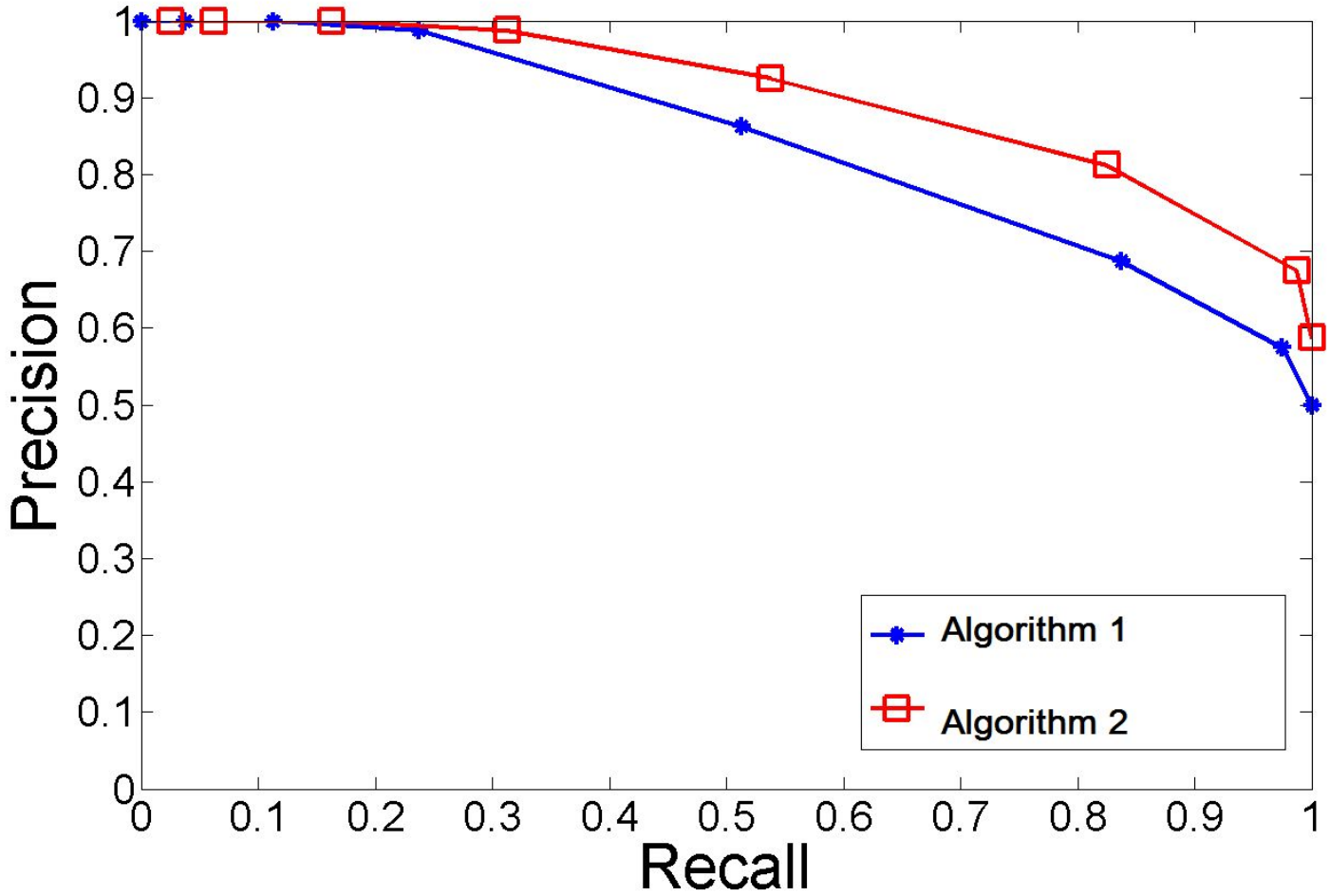
$$F_1 \text{ score} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

- Necesito un balance entre Precision y Recall.



1. Marketing.
2. Fraude en tarjetas.

Métricas Clasificación: Curva PR



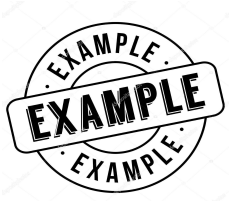
Métricas Clasificación: False Positive Rate (aka Fall-out)

False positive rate (FPR), Fall-out,

probability of false alarm

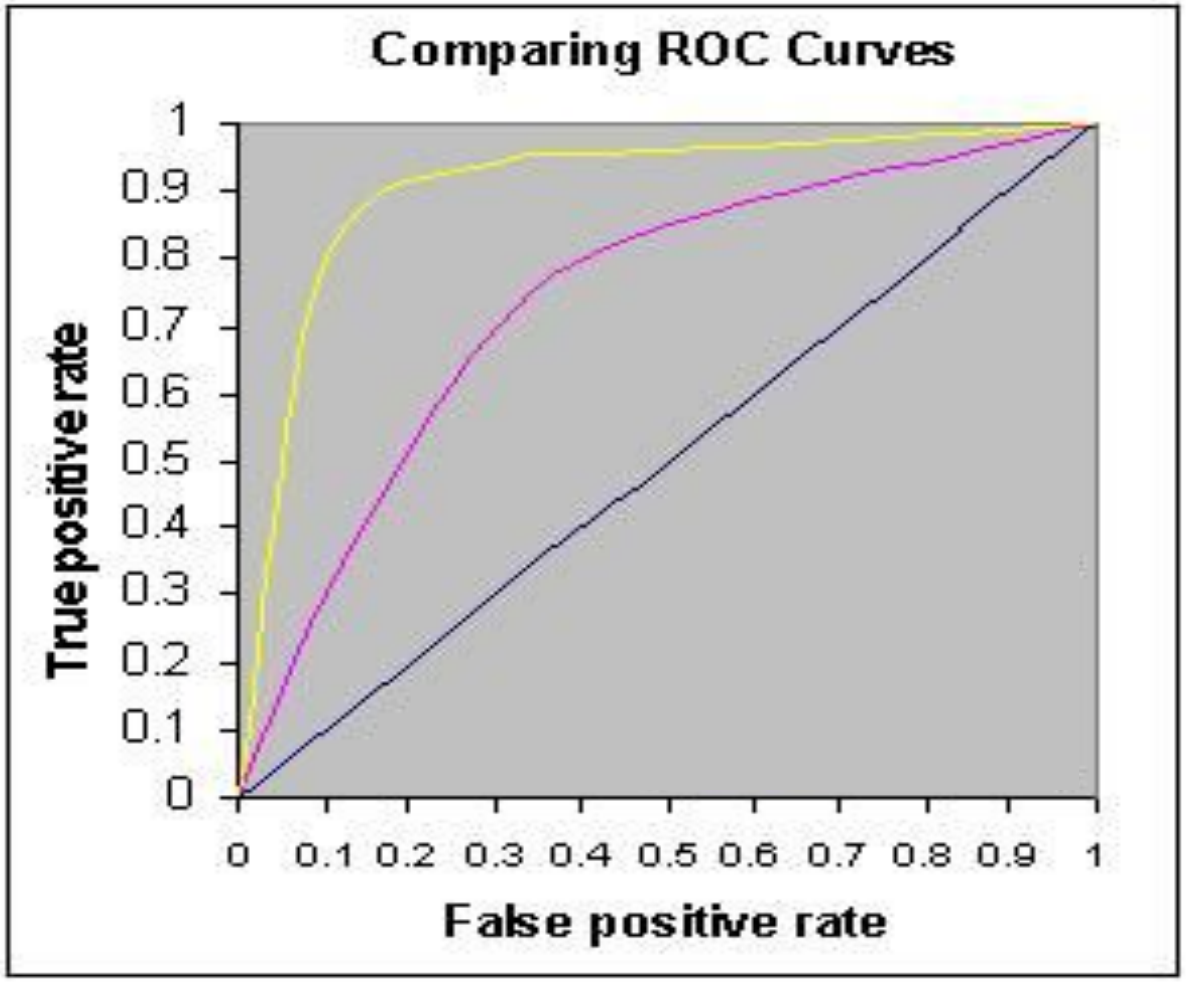
$$= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$$

- No puedo crear falsas alarmas (minimizar)



Epidemias

Métricas Clasificación: Curva ROC



Métricas Clasificación: Area Under the Curve (AUC)

Ranking de modelos: mayor área => mejor modelo

¿Curva PR o curva ROC?

Métricas Clasificación: Area Under the Curve (AUC)

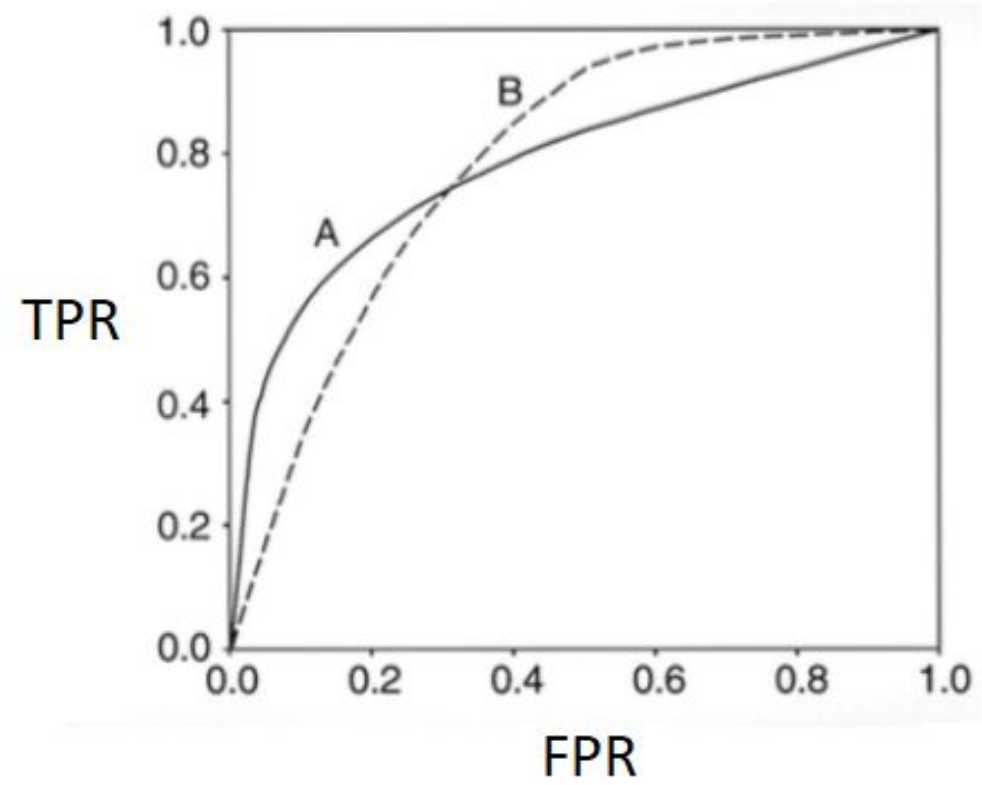
Ranking de modelos: mayor área => mejor modelo

¿Curva PR o curva ROC?

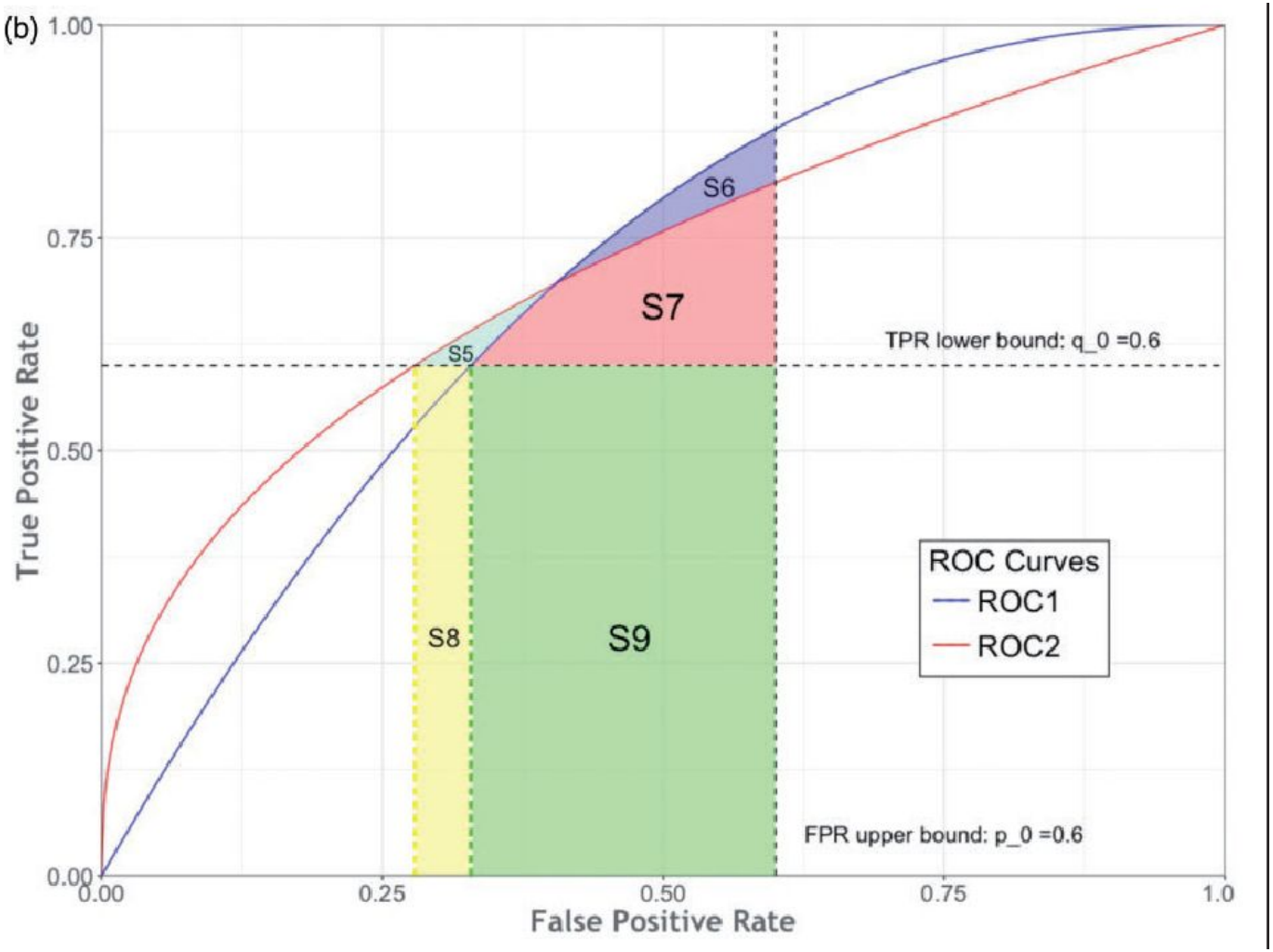
- Curva PR si nos importa la clase positiva y es minoritaria.
- Curva ROC si nos importa la clase positiva y es mayoritaria.
- Curva ROC si ambas clases son igual de importantes.

Métricas Clasificación: Area Under the Curve (AUC)

Ranking de modelos: mayor área => mejor modelo



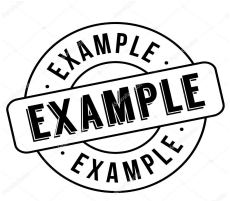
Métricas Clasificación: Two-way partial AUC



Métricas Clasificación: LogLoss

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \cdot log(p_i) + (1 - y_i) \cdot log(1 - p_i)]$$

- Trabajo con probabilidades para cada clase.
- No sólo nos importa la clase predicha, también la seguridad que tenemos.



Cortes asimétricos al asignar clases.

¿Y los sistemas de recomendación cómo se evalúan?

Primero algo de notación…

- *rel(i)*: 1 si el producto ordenado en la posición *i* es relevante. 0 si no lo es.
- *P(i)*: proporción de productos en el top-*i* que son relevantes.
- *n:* nº de productos recomendados por el sistema.
- *m:* nº de productos relevantes que hay en el dataset.
- *Q:* nº de veces que se ejecuta el sistema.

Métricas Ranking: Average Precision at *k* (aka AP@k)

$$AP@k = \frac{1}{min(m,k)} \sum_{i=1}^{min(n,k)} P(i)rel(i)$$

Métricas Ranking: Average Precision at *k* (aka AP@k)

Prediction	Relevance (rel)
1	WRONG (0)
2	RIGHT (1)
3	RIGHT (1)
4	WRONG (0)
5	RIGHT (1)
6	WRONG (0)
7	WRONG (0)
8	WRONG (0)
9	RIGHT (1)
10	WRONG (0)

$$k = 5$$

$$n = nº\ recomendaciones = 5$$

¿AP@5?

Métricas Ranking: Average Precision at *k* (aka AP@k)

Prediction	Relevance (rel)
1	WRONG (0)
2	RIGHT (1)
3	RIGHT (1)
4	WRONG (0)
5	RIGHT (1)
6	WRONG (0)
7	WRONG (0)
8	WRONG (0)
9	RIGHT (1)
10	WRONG (0)

$$AP@k = \frac{1}{min(m,k)} \sum_{i=1}^{min(n,k)} P(i)rel(i)$$

n = 5; k =5; m = 4 ¿AP@5?

P(1) = 0/1; rel(1) = 0

P(2) = 1/2; rel(2) = 1

P(3) = 2/3; rel(3) = 1

P(4) = 2/4; rel(4) = 0

P(5) = 3/5; rel(5) = 1

Métricas Ranking: Normalised Discounted Cumulative Gain at *k* (aka NDCG@k)

Cumulative Gain: Cuanto más relevantes sean los productos recomendados mejor.

$$CG@k = \sum_{i=1}^{k} rel(i)$$

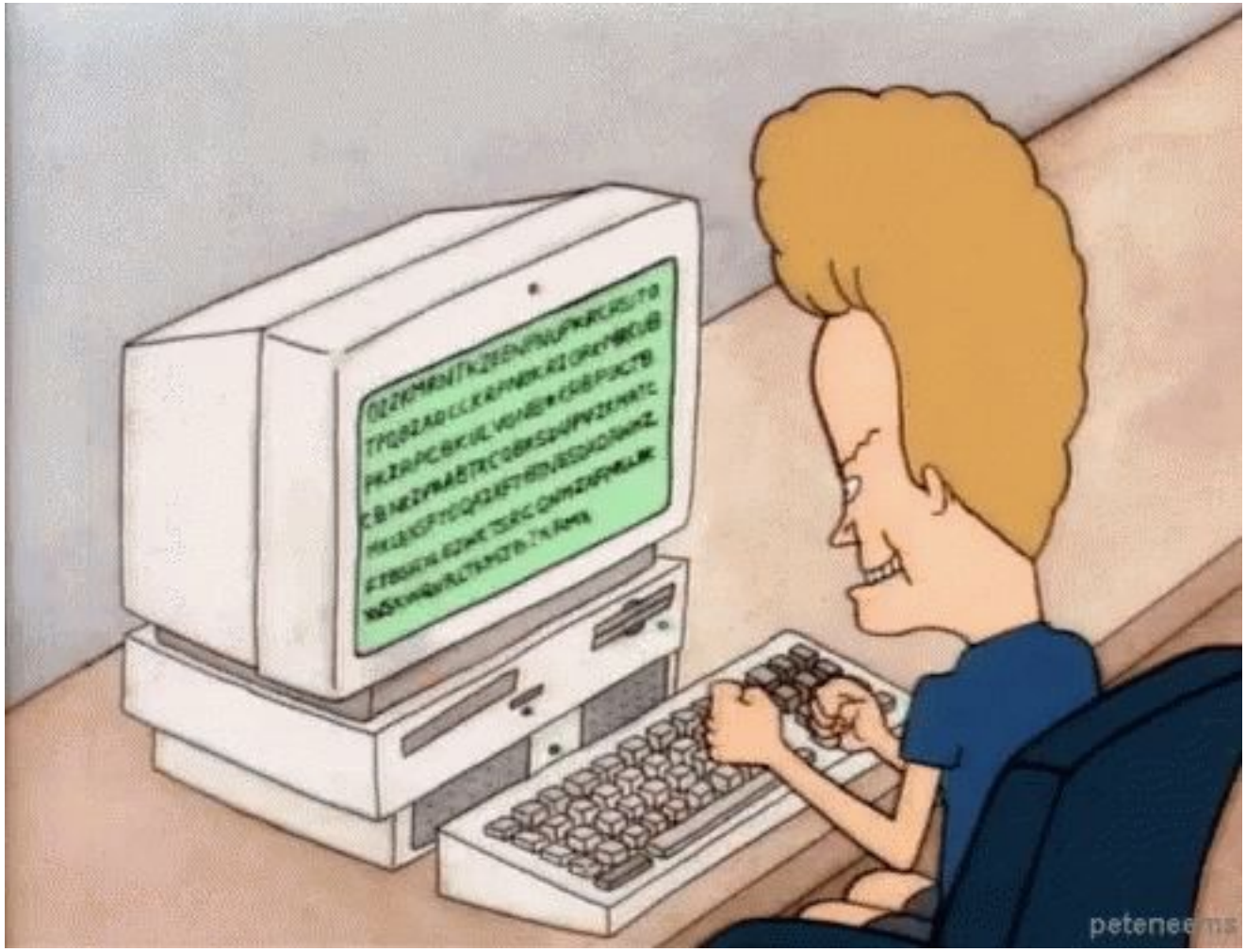
Discounted Cumulative Gain: Cuanto más arriba en el ranking aparezcan los productos más relevantes mejor.

$$DCG@k = \sum_{i=1}^{k} \frac{rel(i)}{log_2(i+1)}$$

Métricas Ranking: Normalised Discounted Cumulative Gain at *k* (aka NDCG@k)

Normalised Discounted Cumulative Gain: Para poder comparar modelos se normaliza dividiendo por el valor ideal de la métrica.

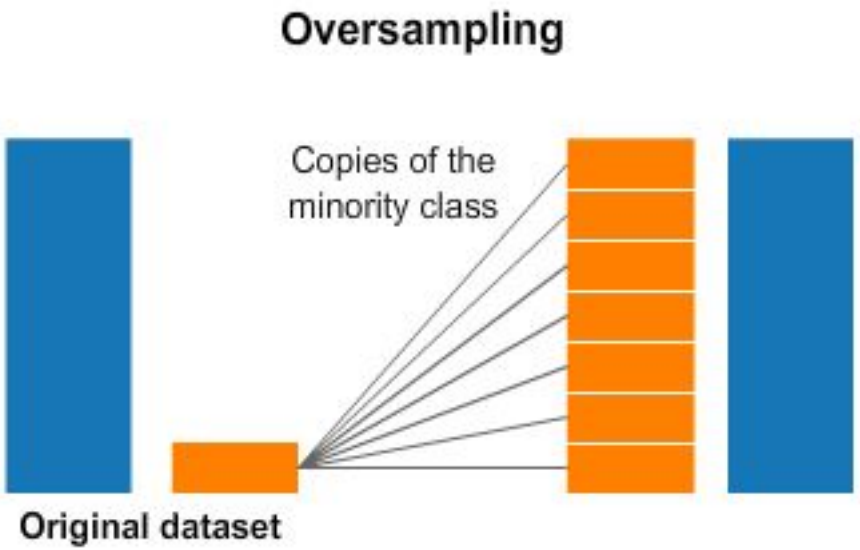
$$NDCG@k = \frac{DCG@k}{IDCG@k}$$



4 | Datasets desbalanceados y otros monstruos.

- Fraude
- Ataques a redes
- Marketing
- Detección enfermedades
- …

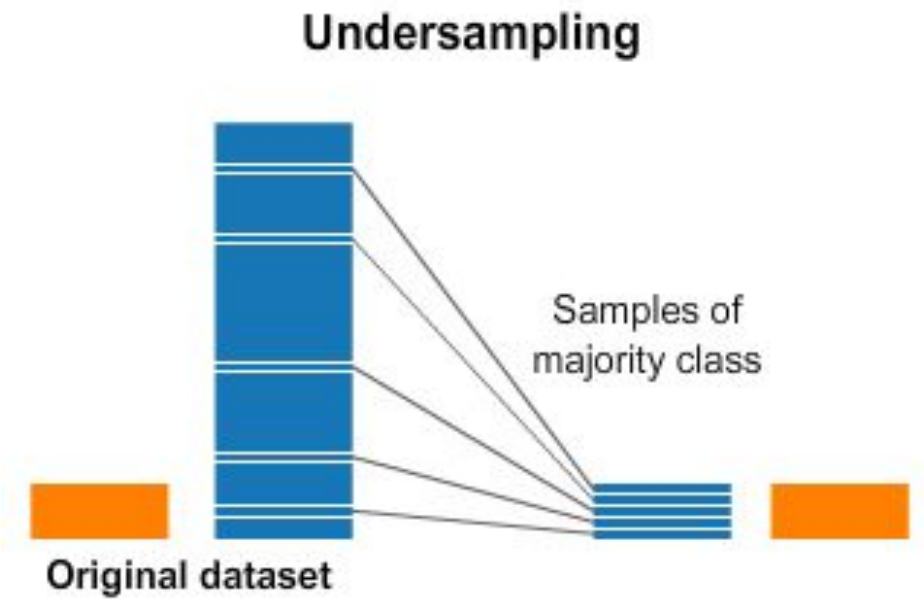
Tratamiento Desbalanceo: Oversampling aleatorio



Overfitting

(source: Kaggle)

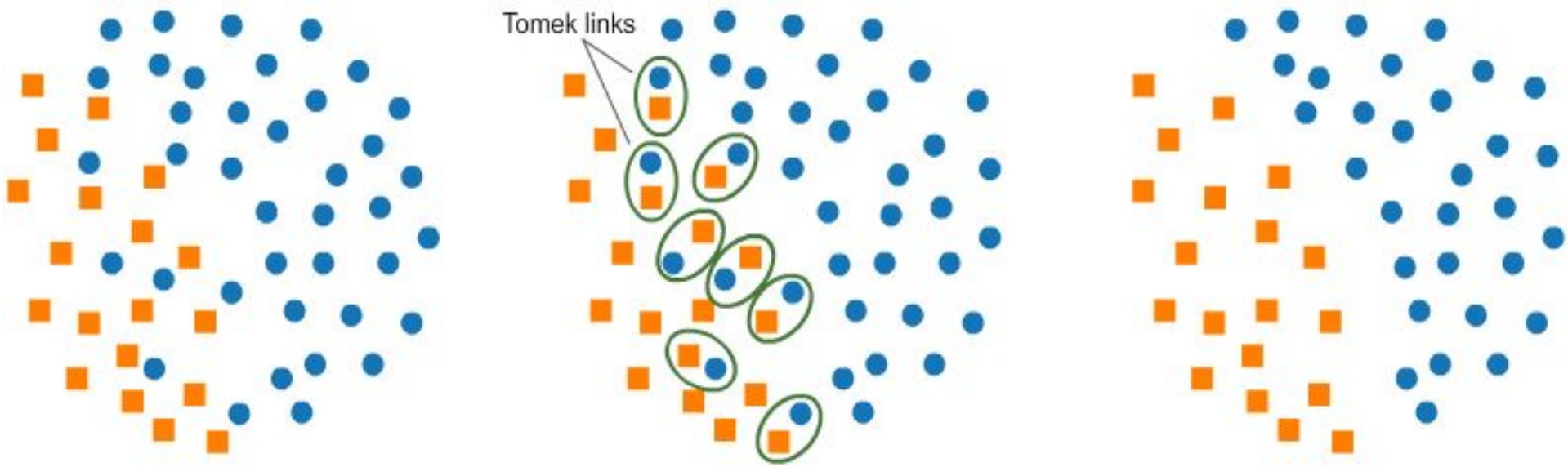
Tratamiento Desbalanceo: Undersampling aleatorio



Perdida de información relevante de la clase mayoritaria.

(source: Kaggle)

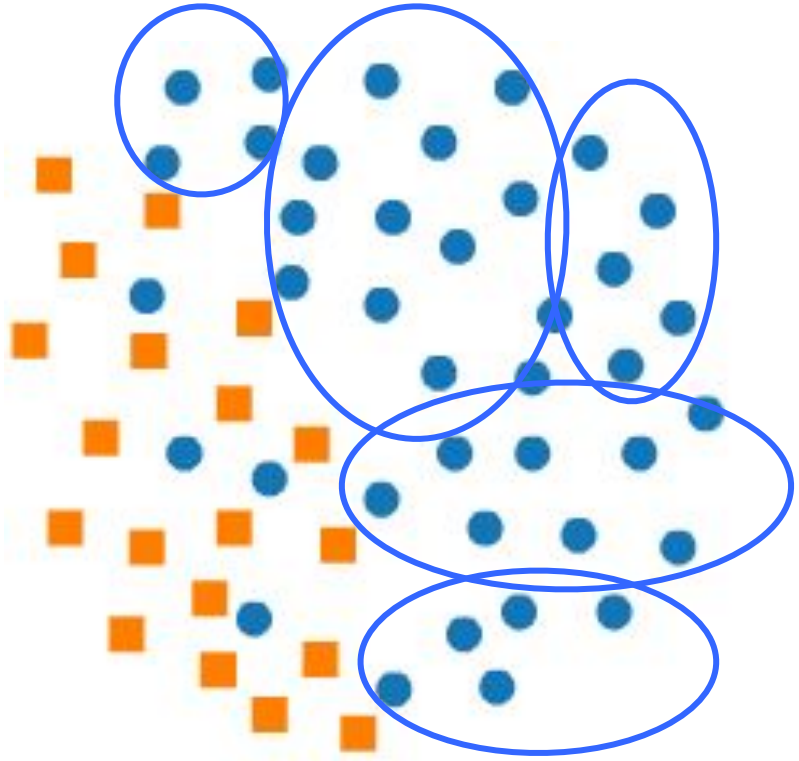
Tratamiento Desbalanceo: Undersampling con enlaces de Tomek



Las fronteras se ensanchan.

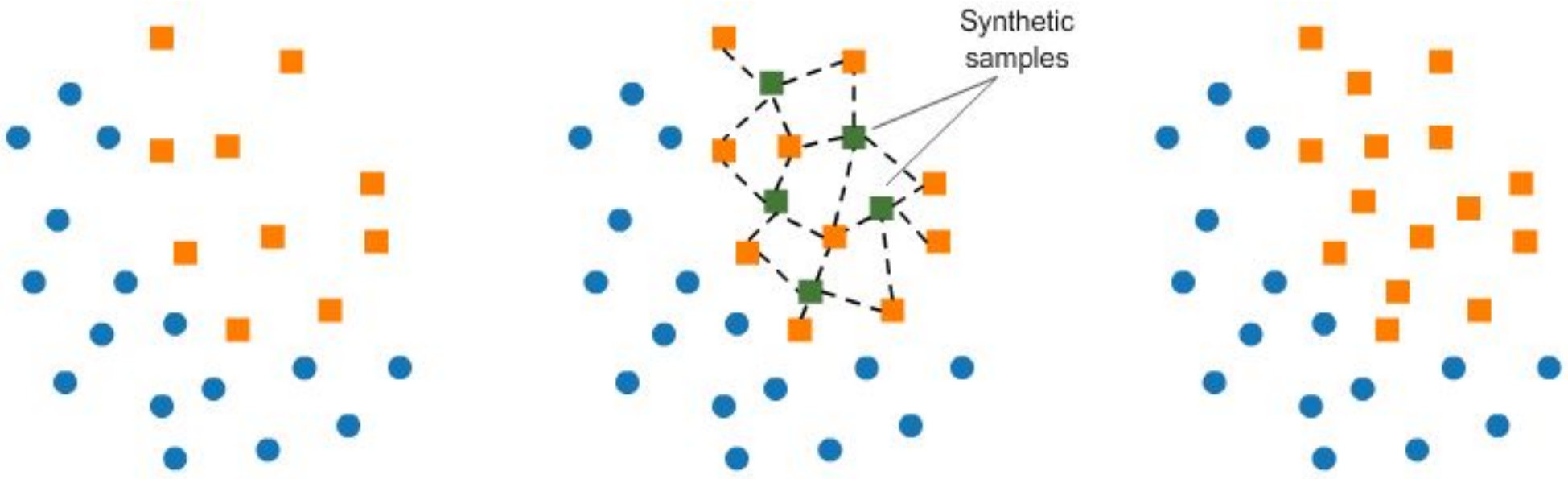
(source: Kaggle)

Tratamiento Desbalanceo: Clustering para undersampling



(source: Kaggle)

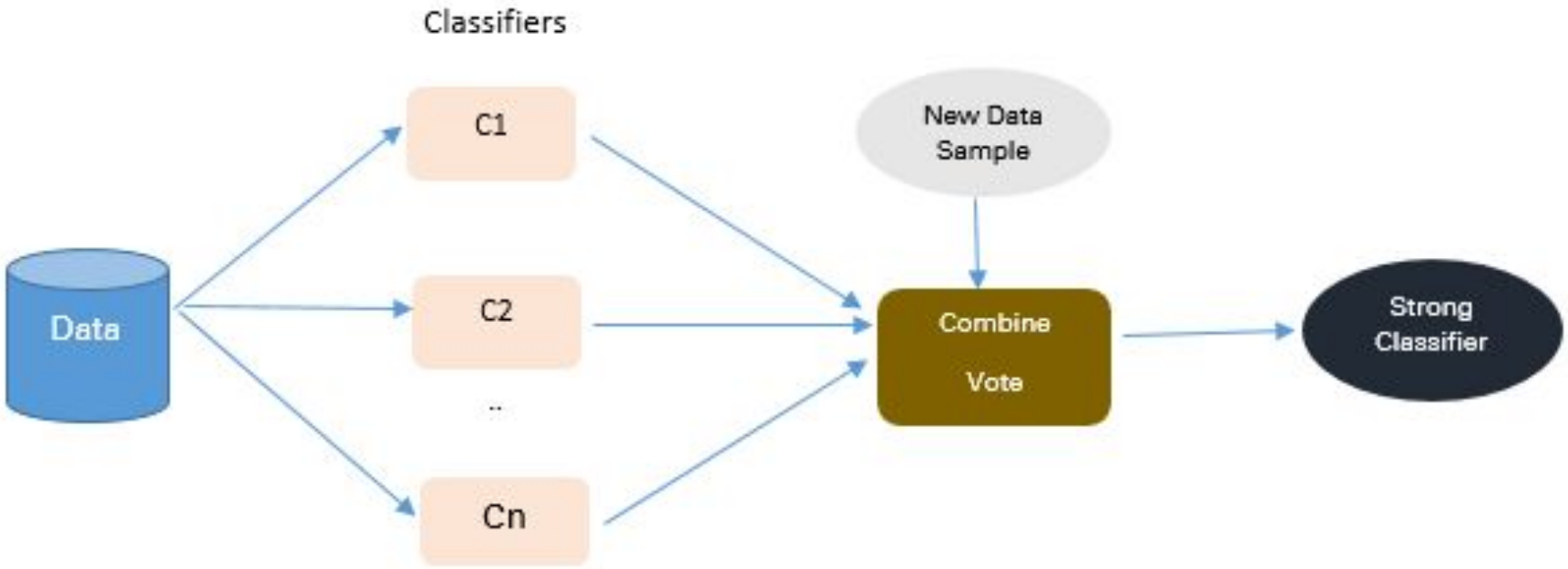
Tratamiento Desbalanceo: Oversampling con SMOTE



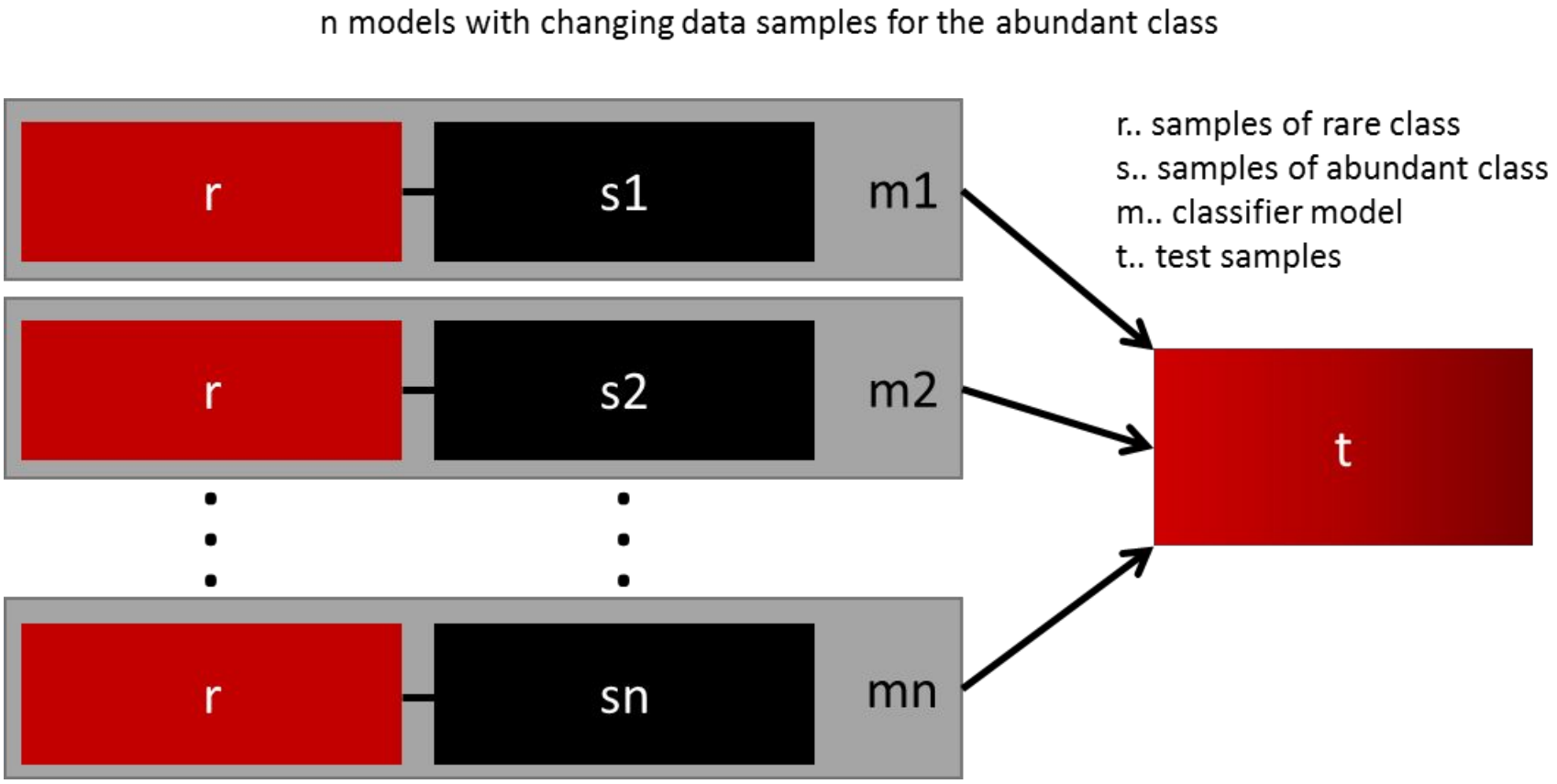
Puntos aislados

(source: Kaggle)

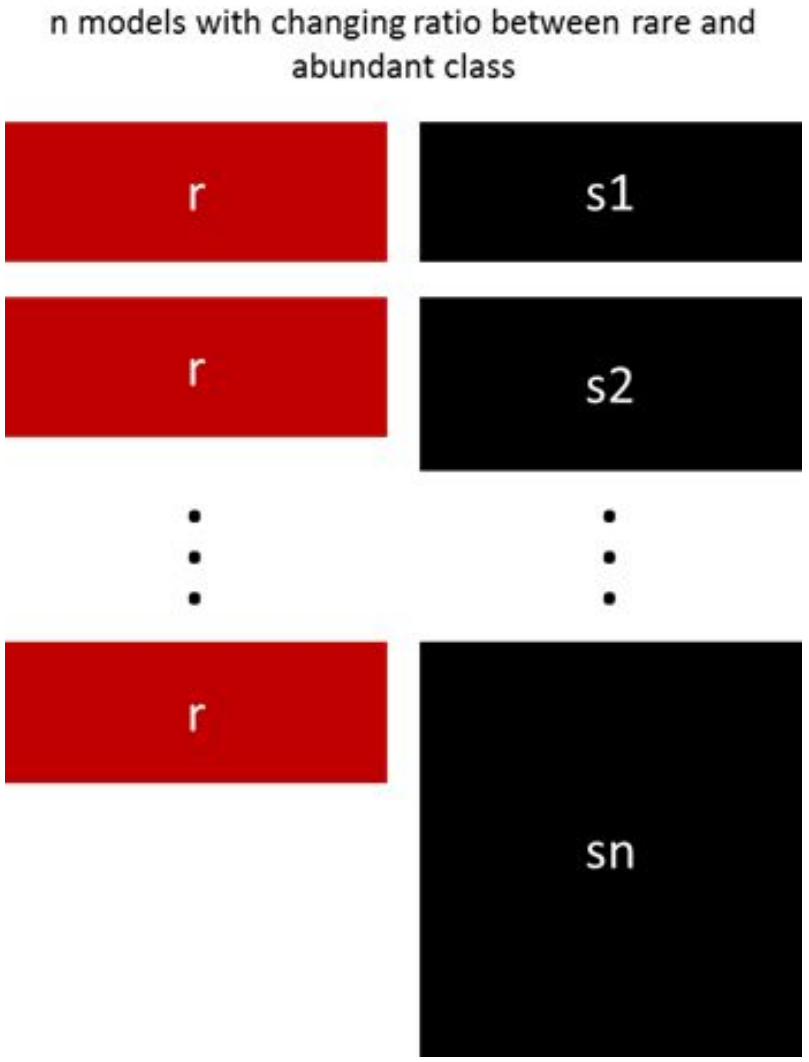
Tratamiento Desbalanceo: Combinación de Modelos

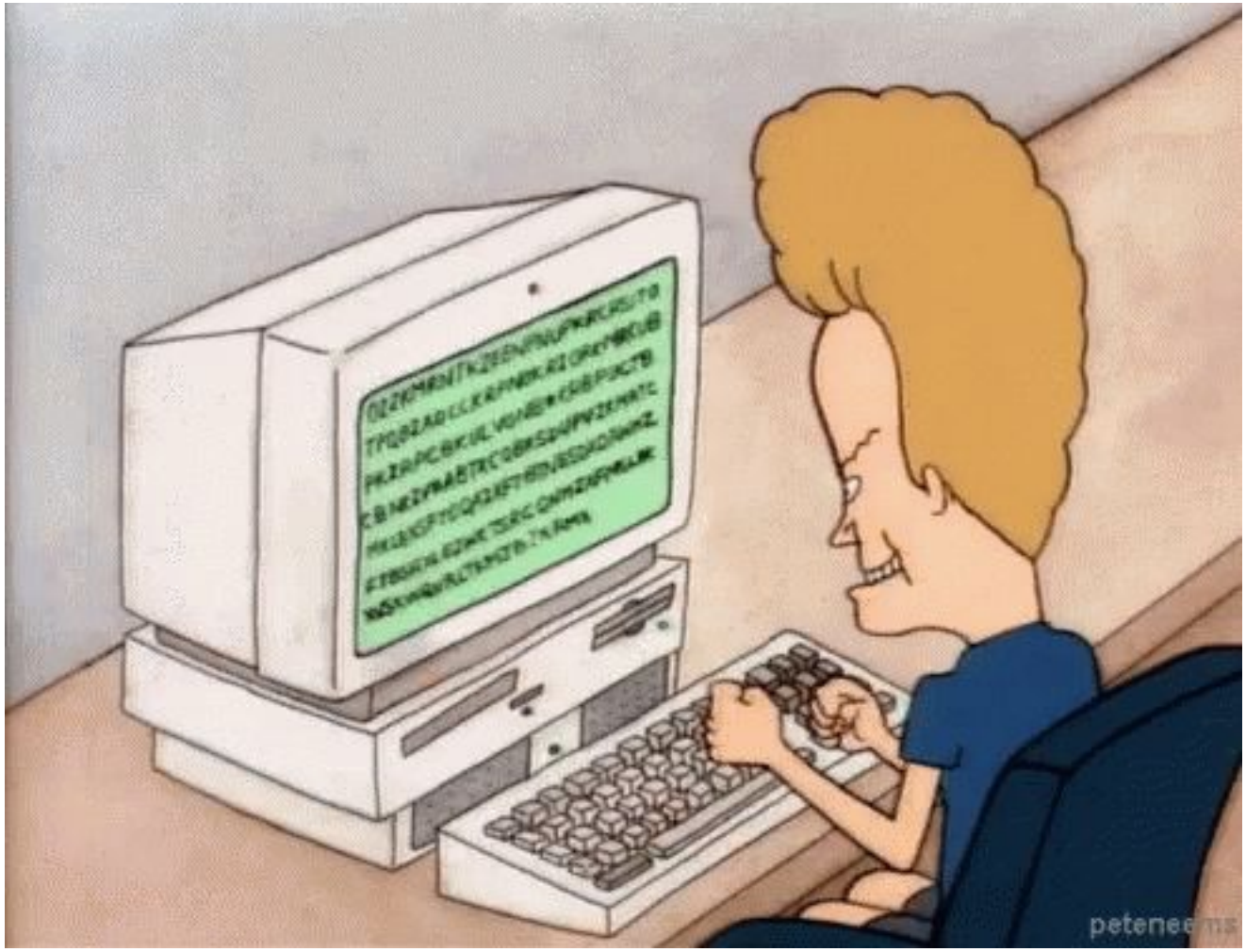


Tratamiento Desbalanceo: Combinación de Modelos

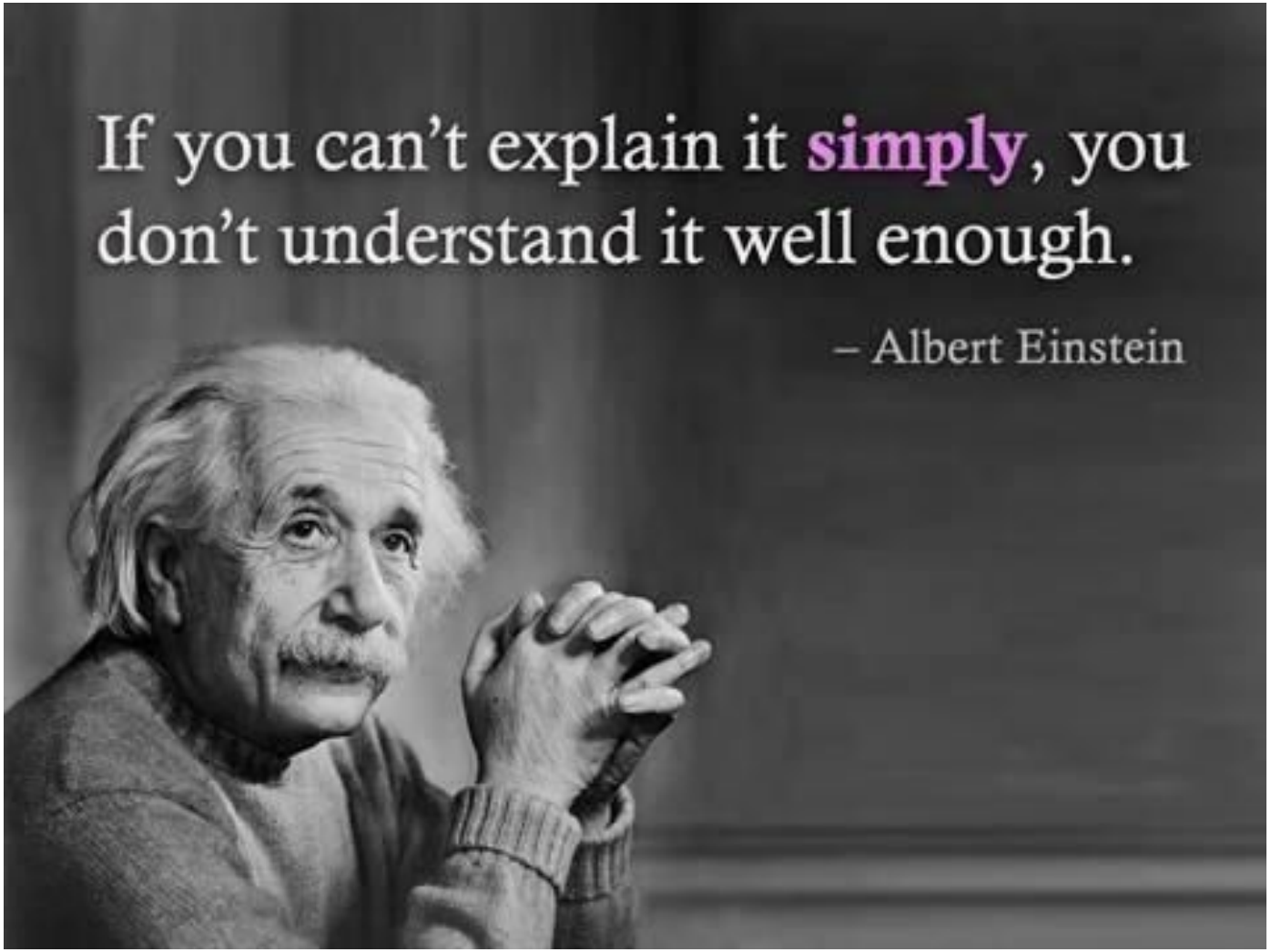


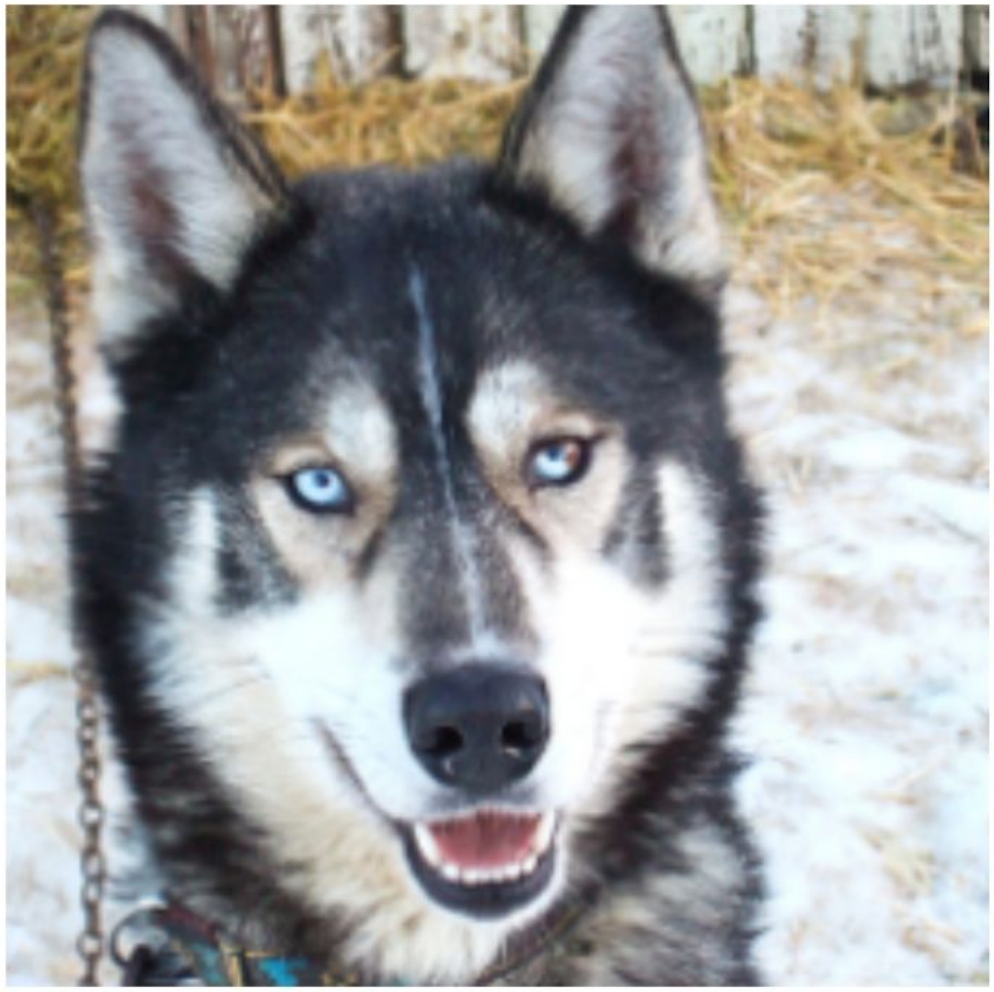
Tratamiento Desbalanceo: Combinación de Modelos



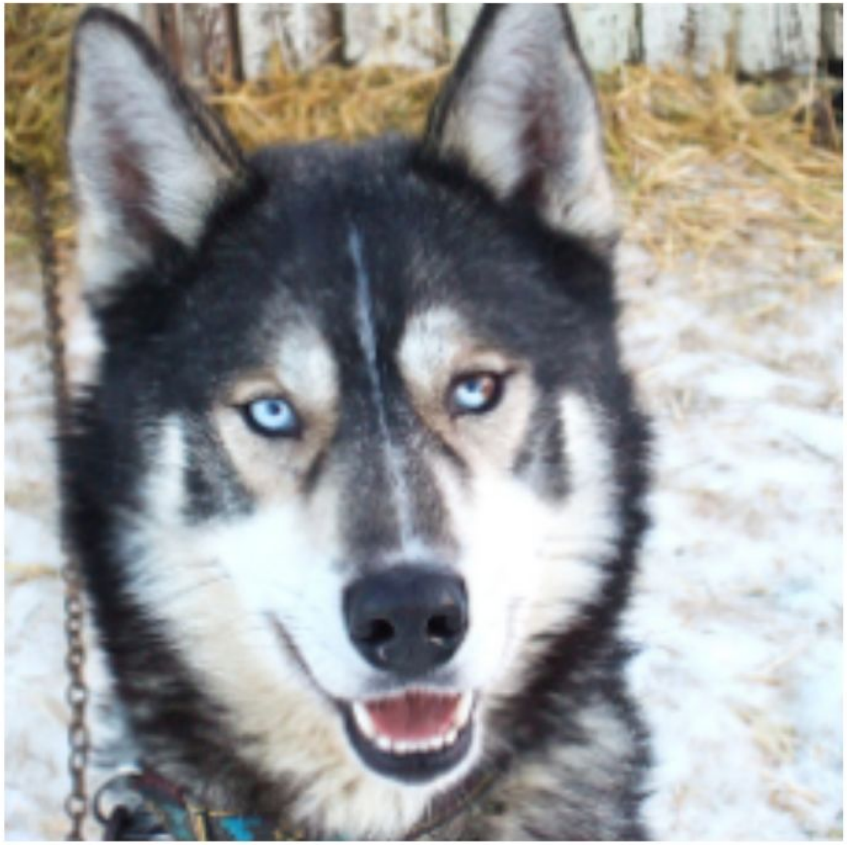


5 Interpretabilidad





(a) Husky classified as wolf



(a) Husky classified as wolf

(b) Explanation

GDPR: Right to Explanation

The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.

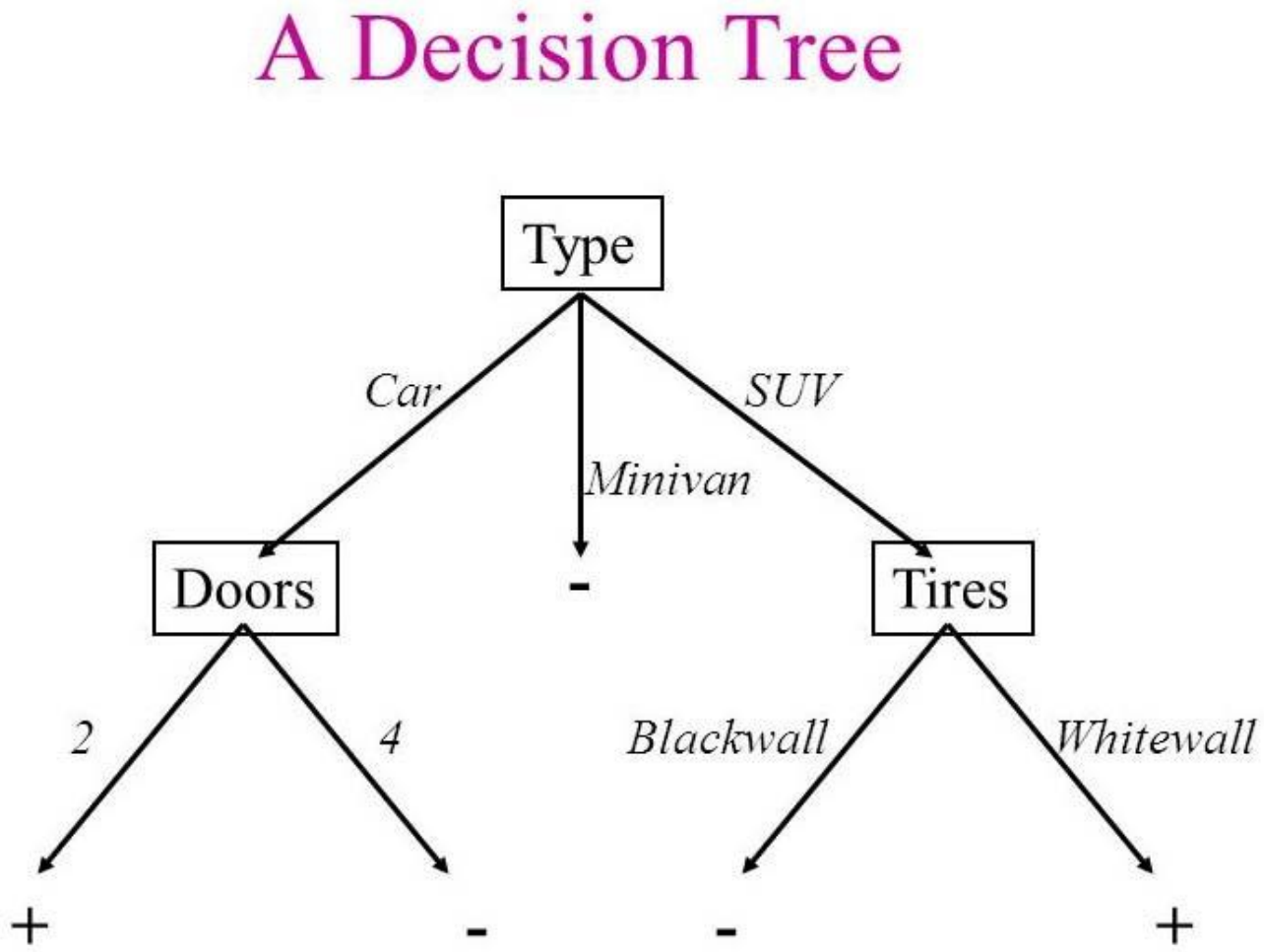
...

In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.

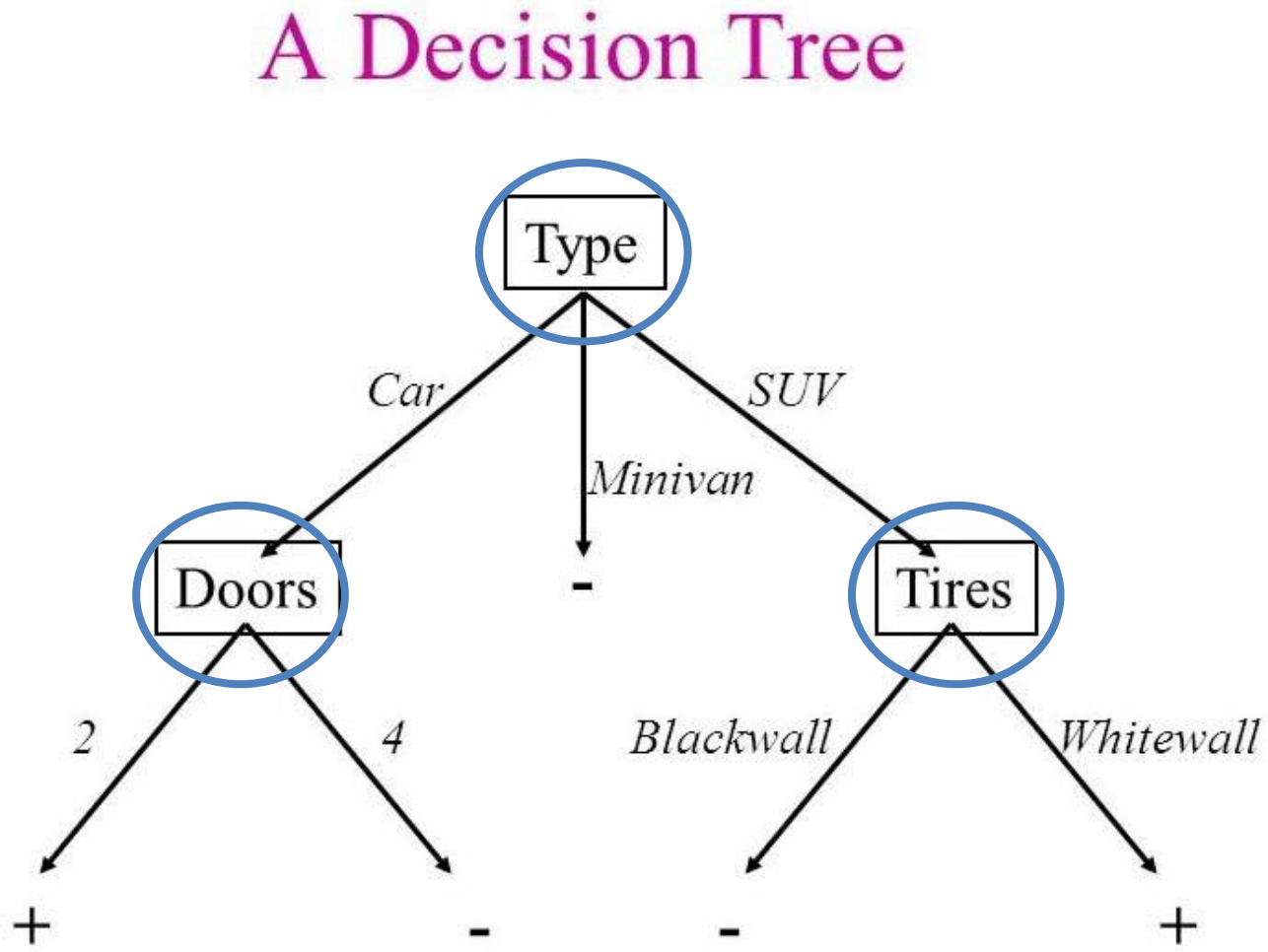
Equal Credit Opportunity Act: Right to Explanation

(2) Statement of specific reasons. The statement of reasons for adverse action required by paragraph (a)(2)(i) of this section must be specific and indicate the principal reason(s) for the adverse action. Statements that the adverse action was based on the creditor's internal standards or policies or that the applicant, joint applicant, or similar party failed to achieve a qualifying score on the creditor's credit scoring system are insufficient.

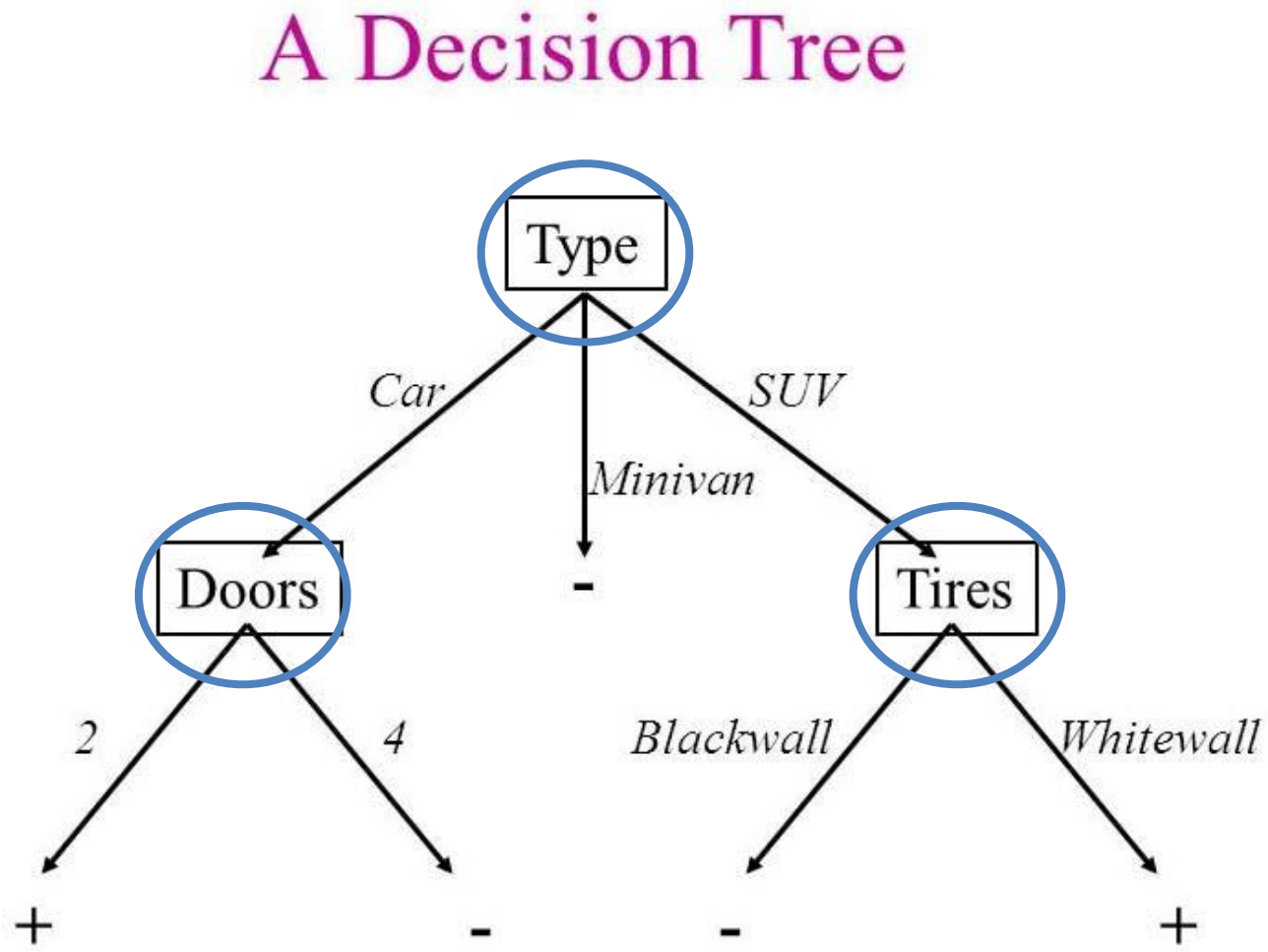
Interpretabilidad local y global



Interpretabilidad global

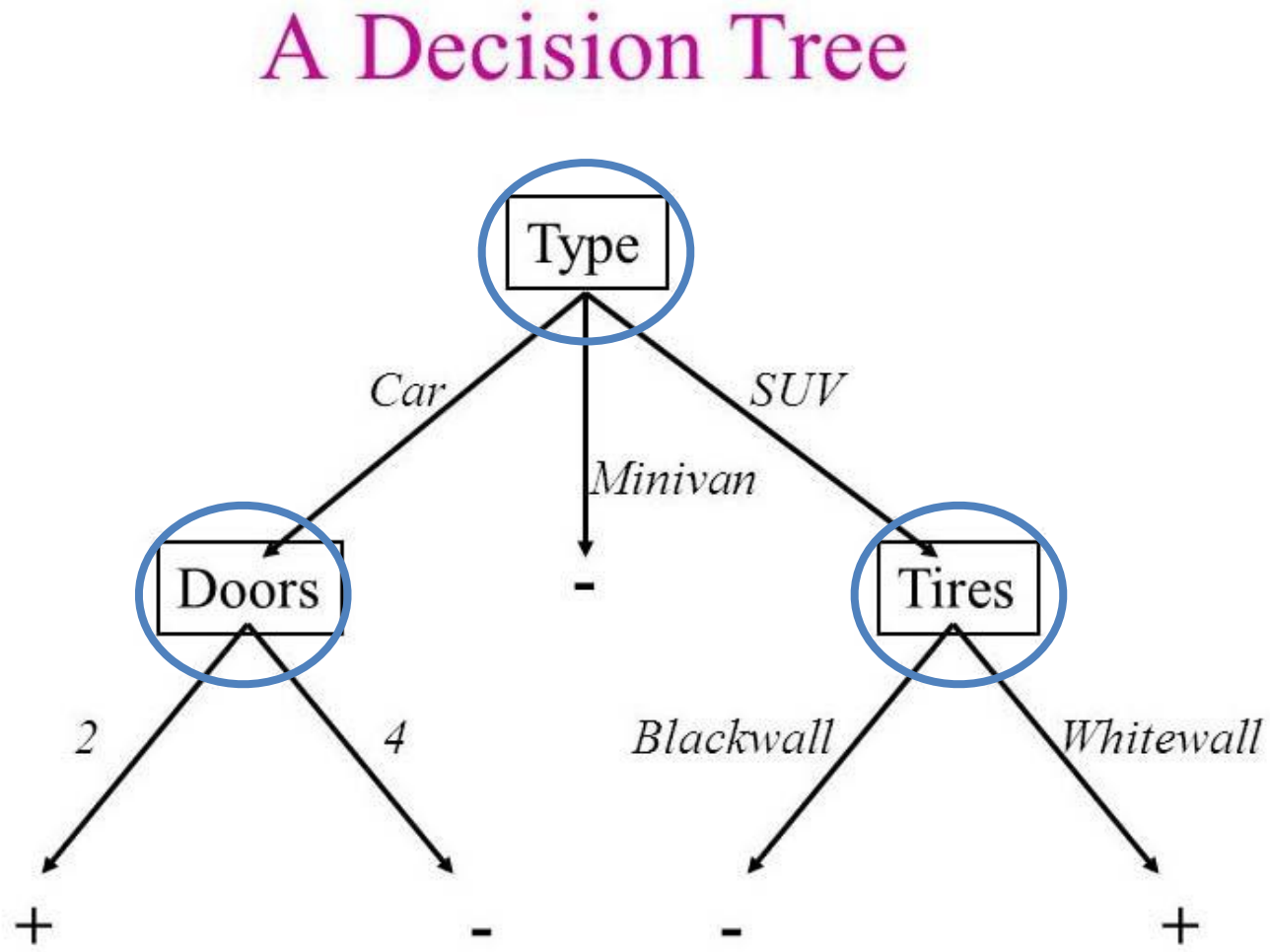


Interpretabilidad global



Type > Doors
Type > Tires

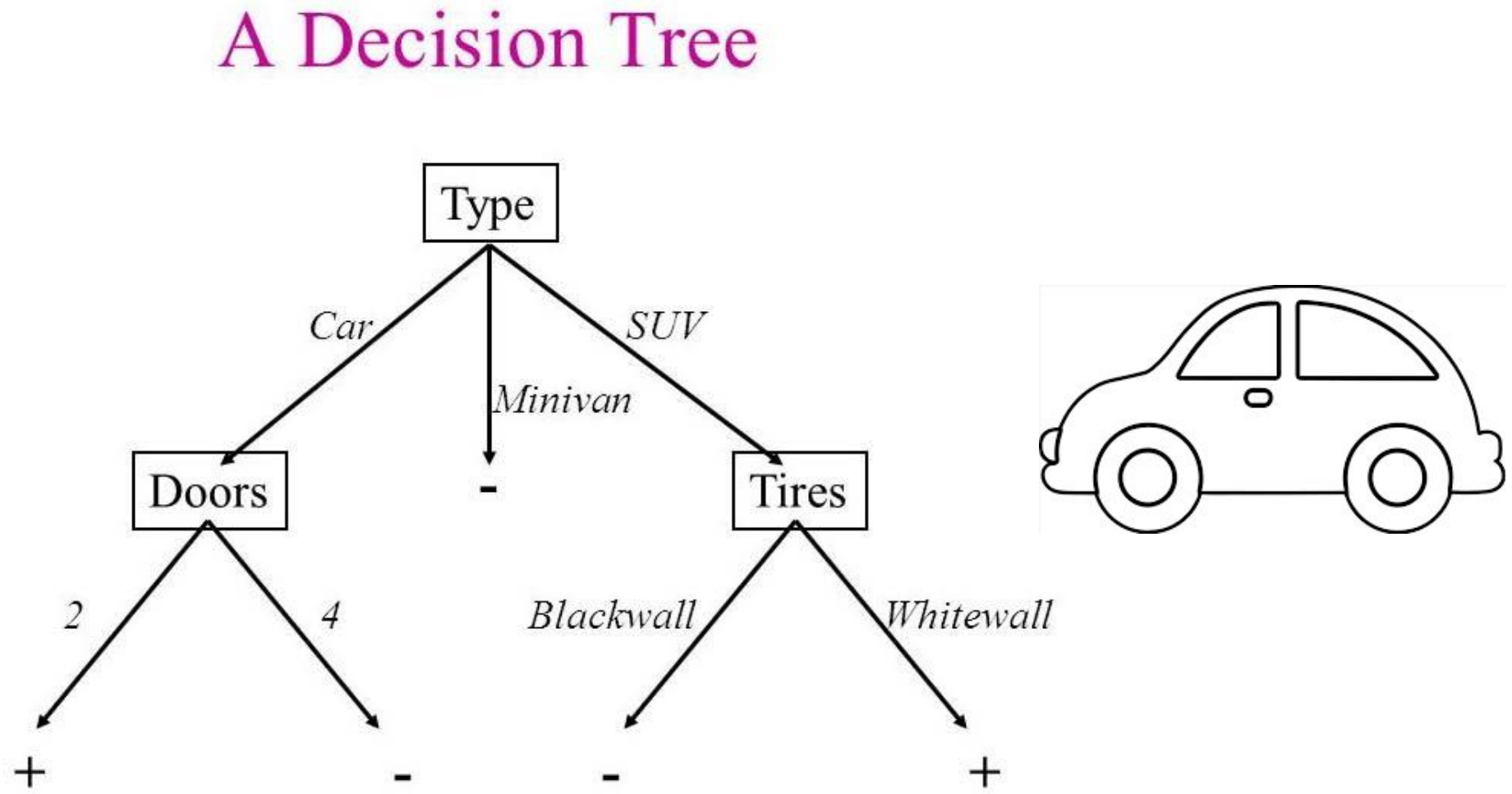
Interpretabilidad global



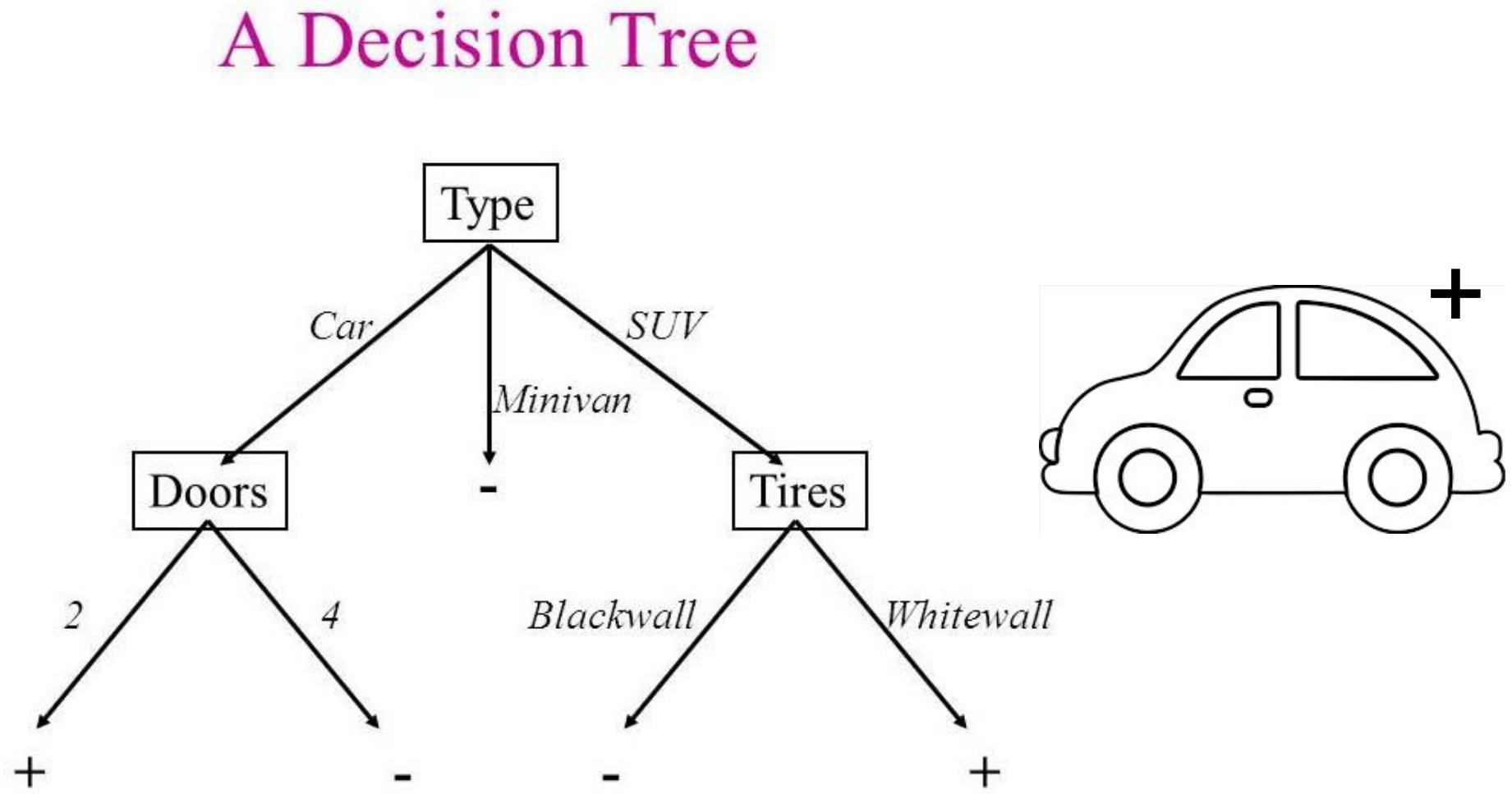
Type > Doors
Type > Tires

¿Doors > Tires?
¿Tires > Doors?
¿Doors = Tires?

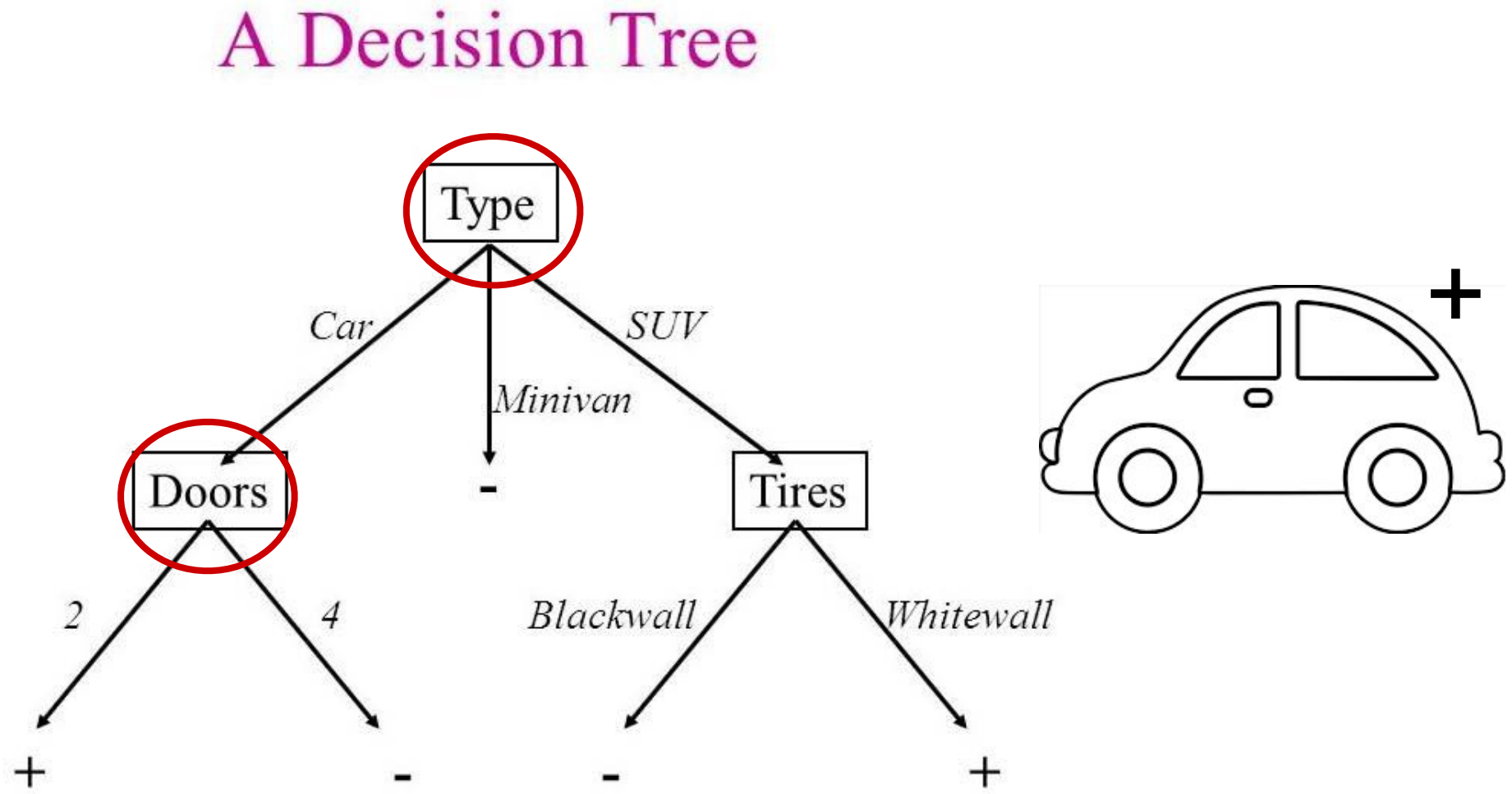
Interpretabilidad local



Interpretabilidad local



Interpretabilidad local



Ejemplo de librería de interpretabilidad: LIME

Local Interpretable Model-agnostic Explanations

LIME

1. Construir el modelo de caja negra de la forma habitual.

LIME

1. Construir el modelo de caja negra de la forma habitual.
2. Aplicarlo al conjunto de test y elegir el punto cuya predicción queremos explicar.

LIME

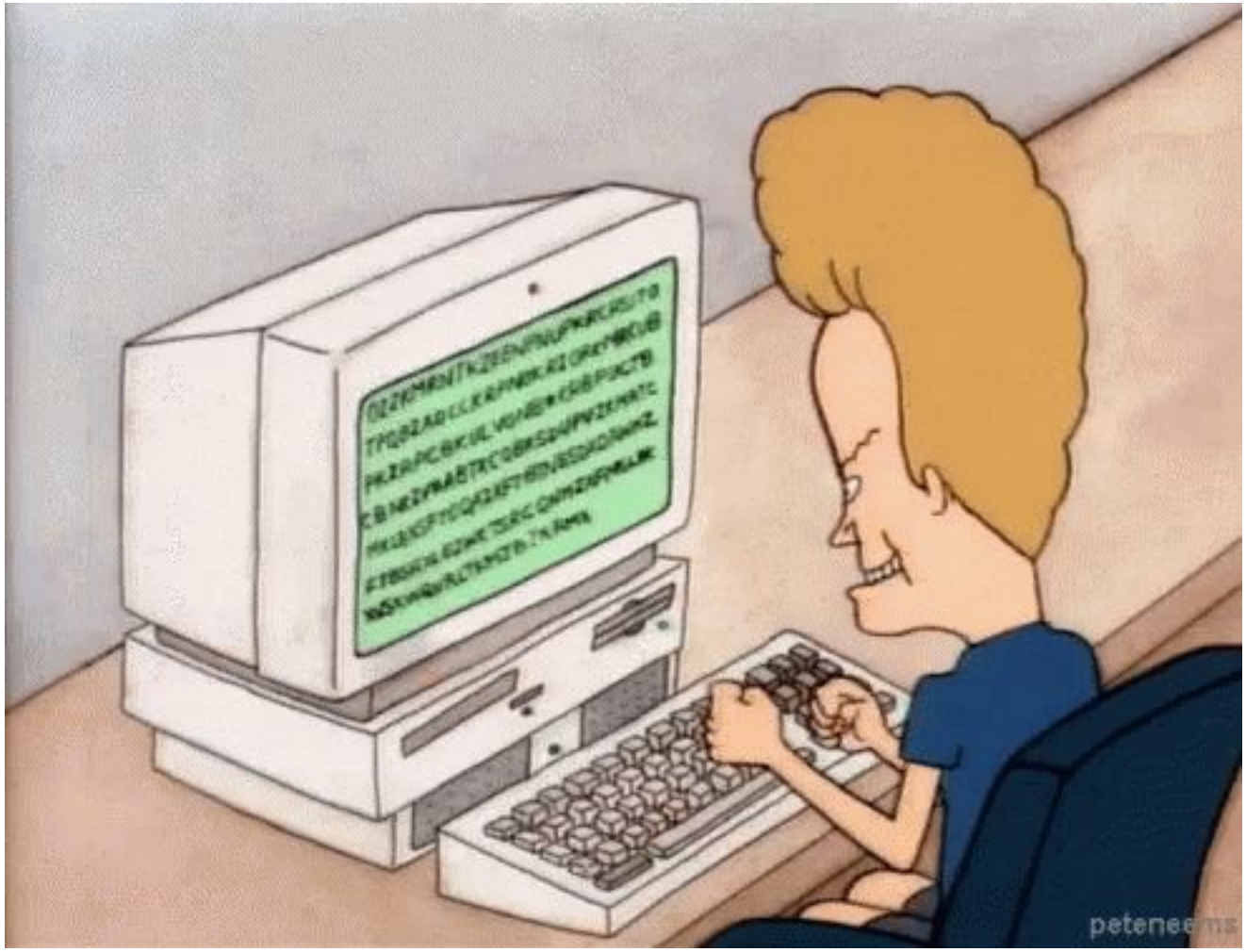
1. Construir el modelo de caja negra de la forma habitual.
2. Aplicarlo al conjunto de test y elegir el punto cuya predicción queremos explicar.
3. LIME genera puntos sintéticos en el espacio, idealmente cercanos al punto elegido, y aplica sobre ellos el modelo.

LIME

1. Construir el modelo de caja negra de la forma habitual.
2. Aplicarlo al conjunto de test y elegir el punto cuya predicción queremos explicar.
3. LIME genera puntos sintéticos en el espacio, idealmente cercanos al punto elegido, y aplica sobre ellos el modelo.
4. LIME ajusta un modelo interpretable (*e.g.* Lasso, Ridge Regression…) sobre el conjunto de puntos sintéticos.

LIME

1. Construir el modelo de caja negra de la forma habitual.
2. Aplicarlo al conjunto de test y elegir el punto cuya predicción queremos explicar.
3. LIME genera puntos sintéticos en el espacio, idealmente cercanos al punto elegido, y aplica sobre ellos el modelo.
4. LIME ajusta un modelo interpretable (*e.g.* Lasso, Ridge Regression…) sobre el conjunto de puntos sintéticos.
5. Para obtener la importancia de cada atributo LIME pondera los puntos según su cercanía al punto objetivo.



6 Antes de terminar…

Afi Escuela de Finanzas