Sistemas Recomendadores IIC-3633

Explicabilidad en Sistemas Recomendadores

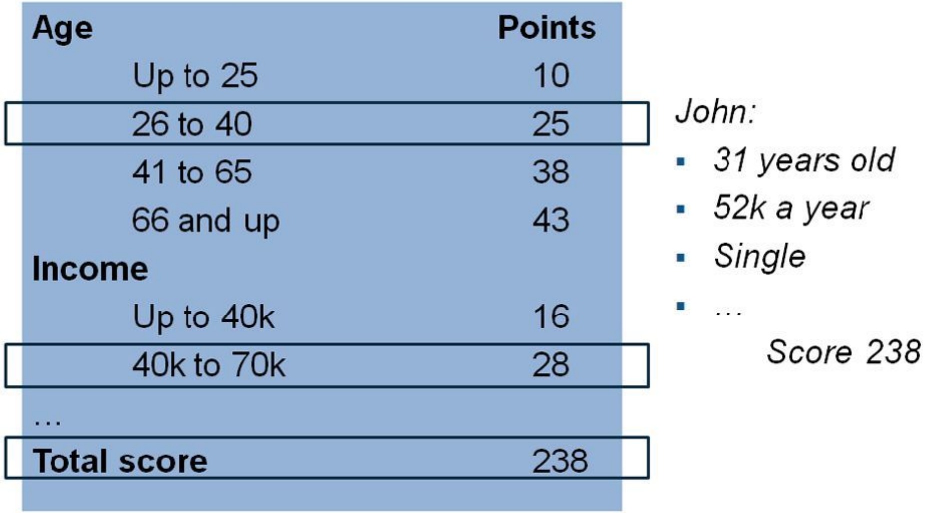
Esta clase

1. Explainable AI
2. Explicabilidad en sistemas de recomendación

Motivación de XAI

Salud / Tamizaje

Predicción de a quien dar un crédito bancario y por qué?



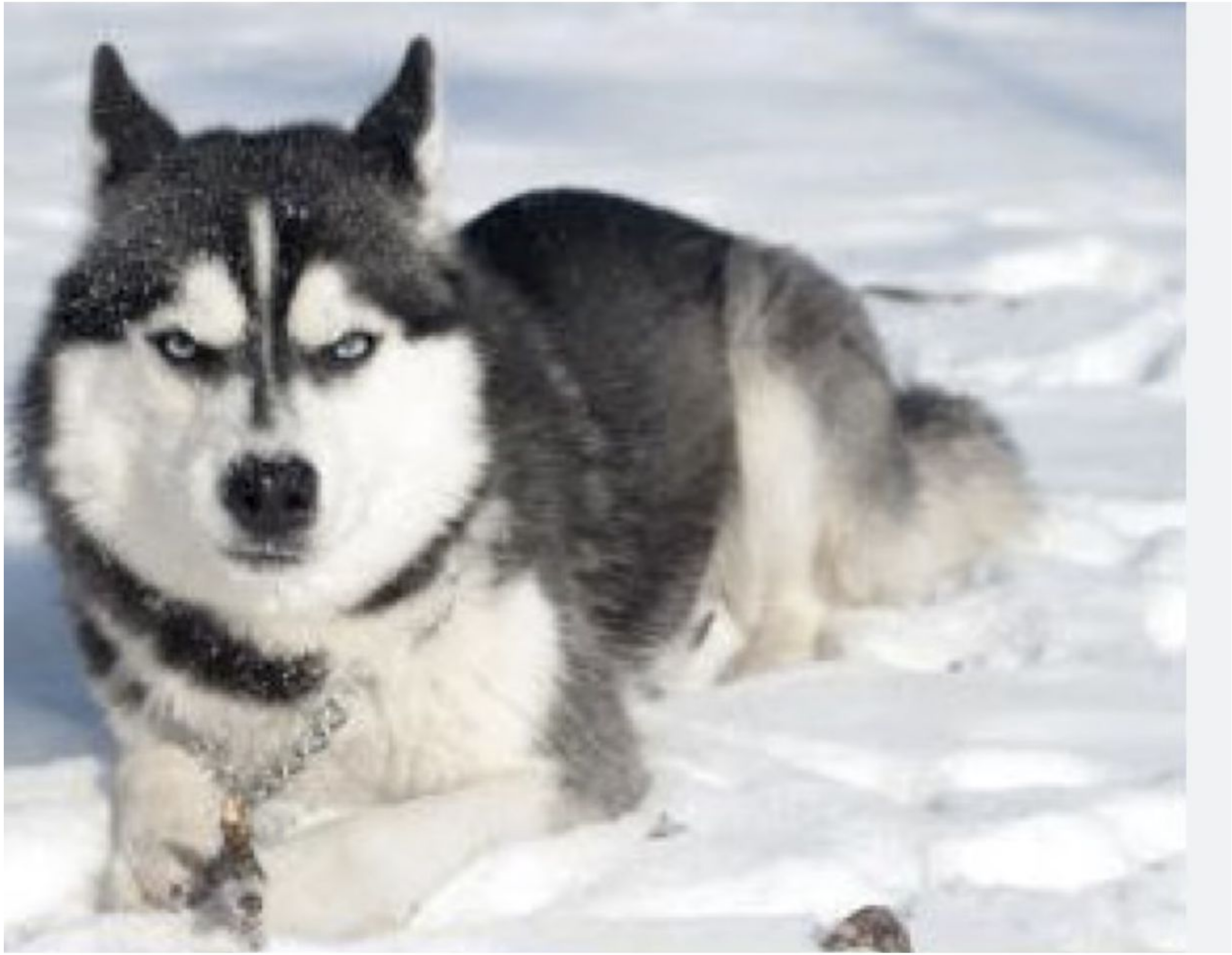
Age	Points
Up to 25	10
26 to 40	25
41 to 65	38
66 and up	43
Income	
Up to 40k	16
40k to 70k	28
…	
Total score	238

John:

- *31 years old*
- *52k a year*
- *Single*
- …

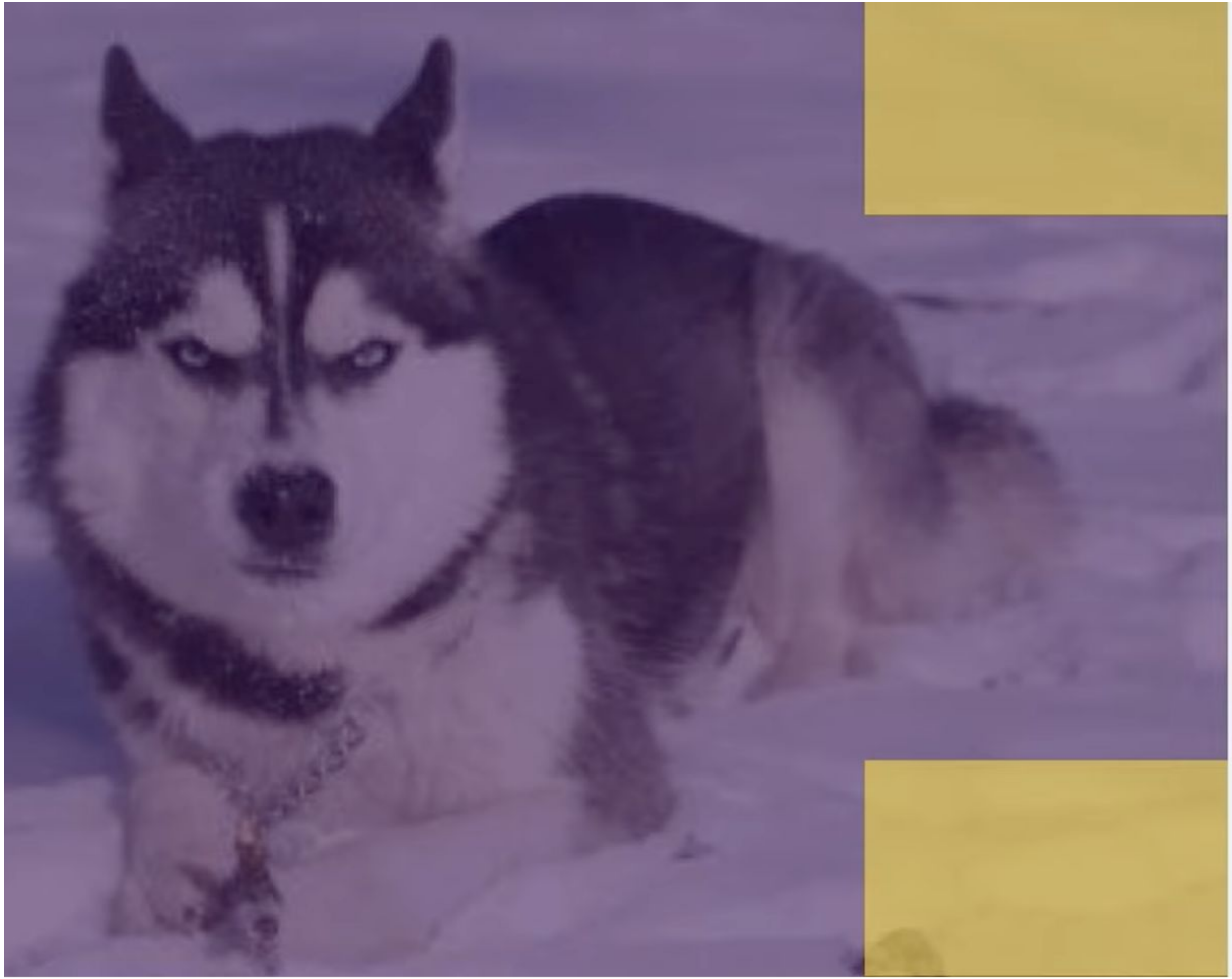
Score 238

Husky o
Lobo?

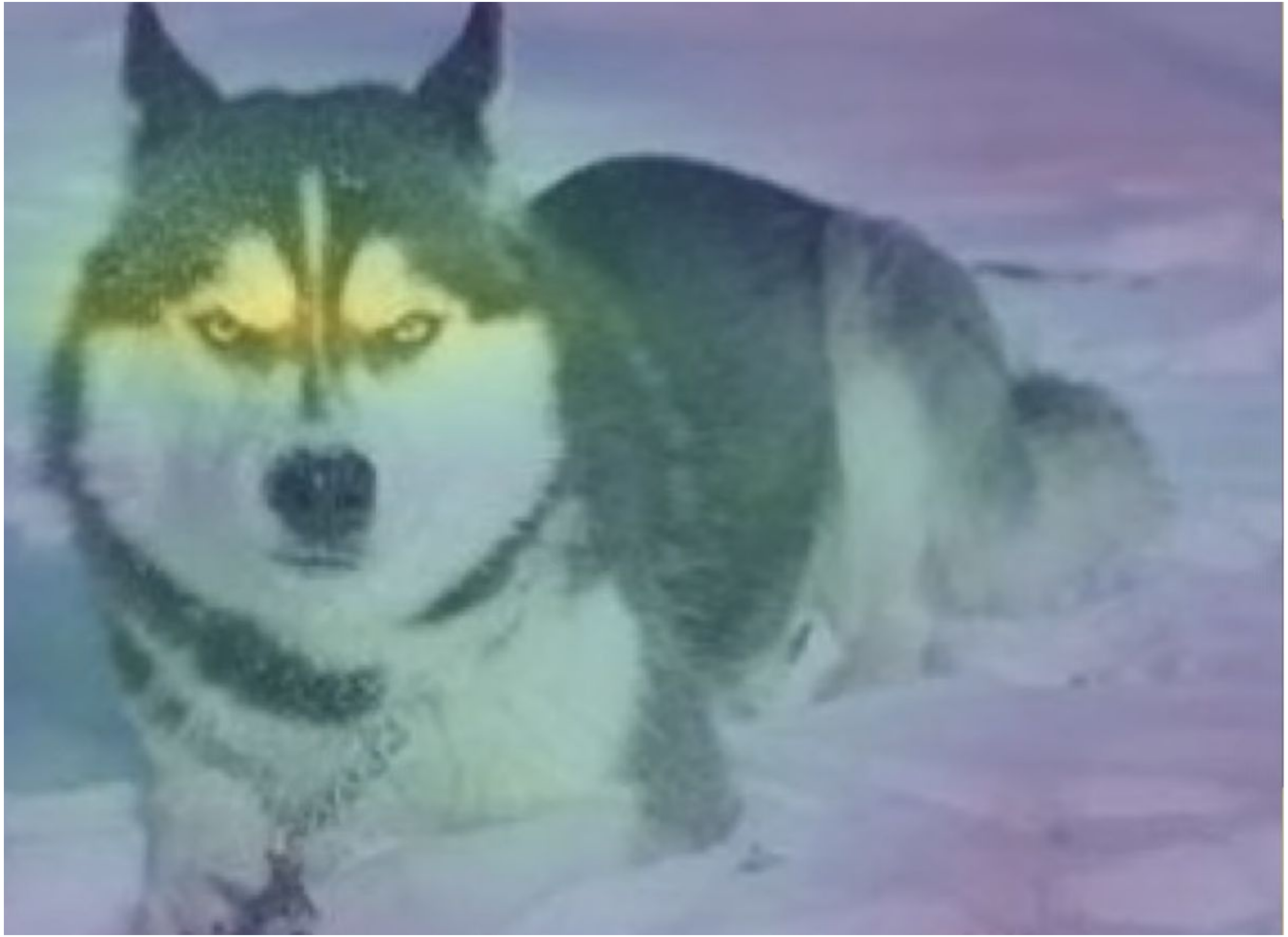


Área de la imagen en la que se fijó el modelo para clasificarlo como lobo.

Toma la decisión correcta pero basado en malas razones (la nieve)

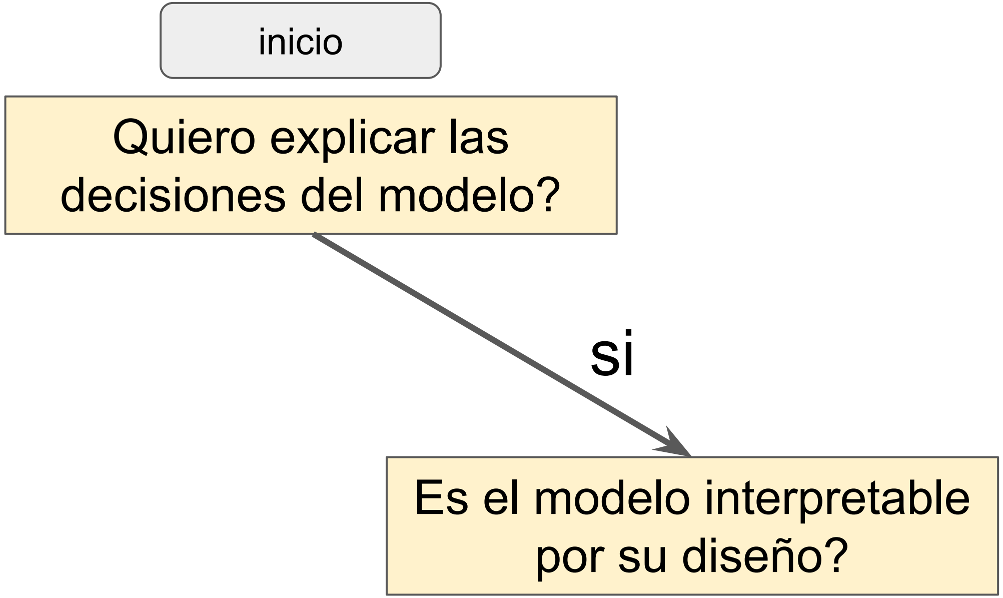


El modelo debería prestar atención a esta zona para determinar si es un lobo o no.



inicio

Quiero explicar las decisiones del modelo?

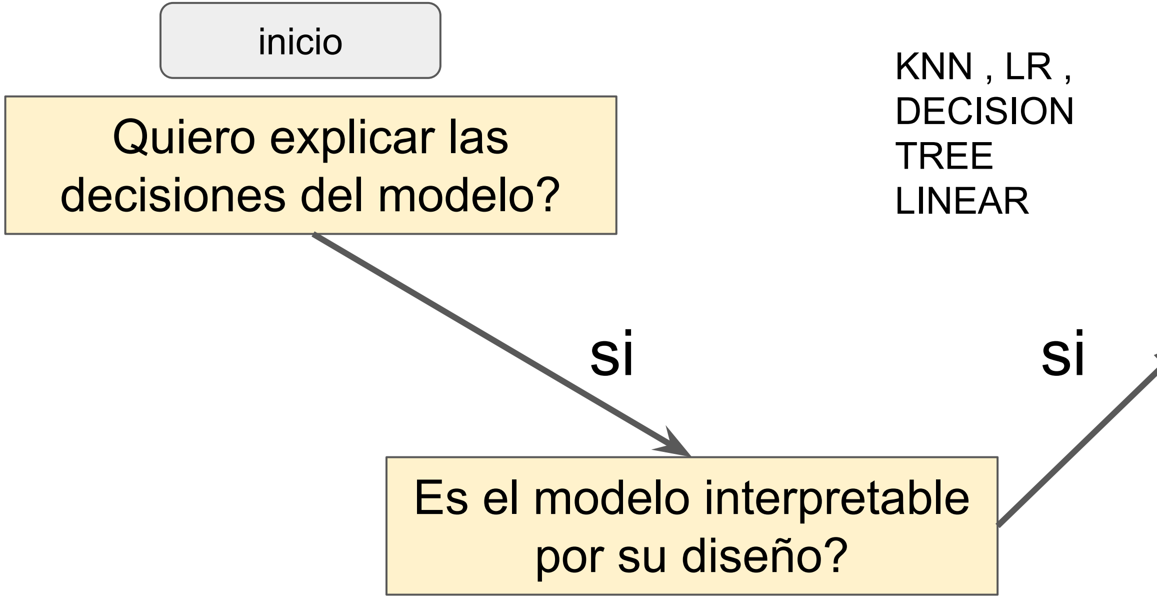


inicio

Quiero explicar las decisiones del modelo?

si

Es el modelo interpretable por su diseño?



inicio

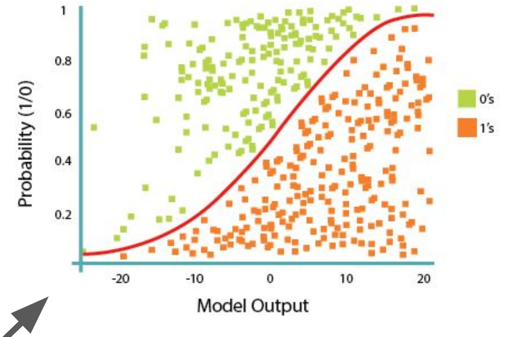
Quiero explicar las decisiones del modelo?

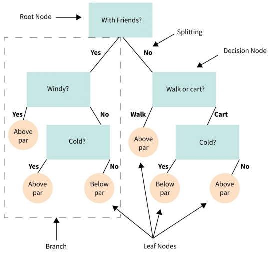
si

Es el modelo interpretable por su diseño?

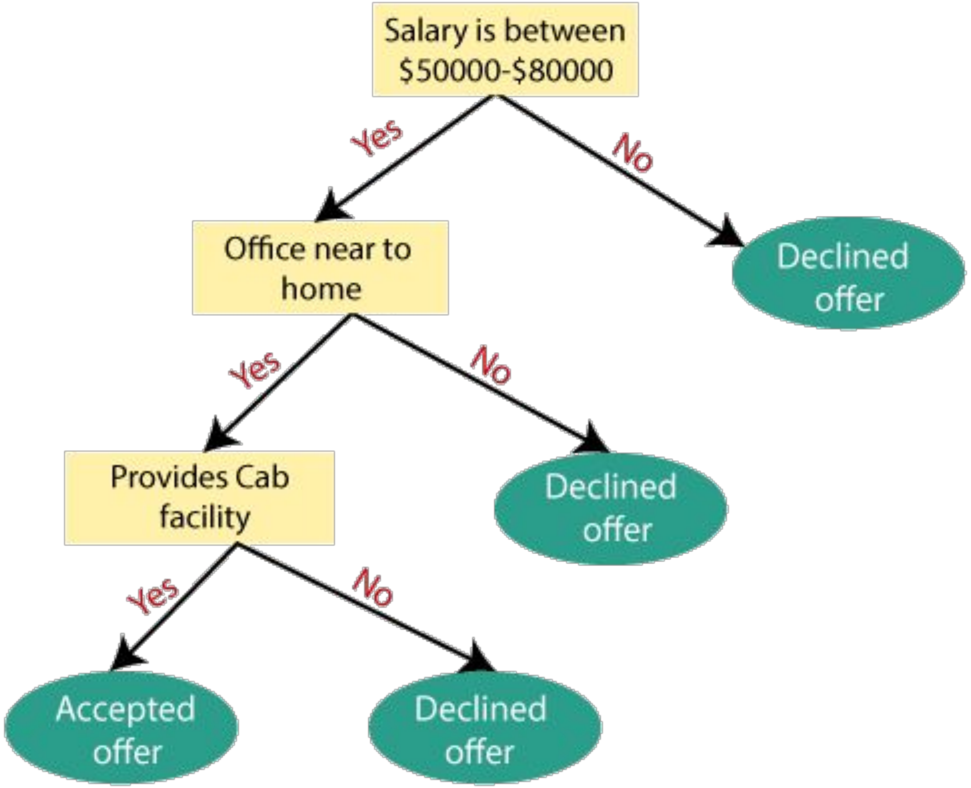
si

KNN , LR , DECISION TREE LINEAR





Muy bueno para hacerse una idea de los features más importantes a la hora de hacer inferencia !!


Salary is between $50000-$80000
Yes
No
Office near to home
Declined offer
Yes
No
Provides Cab facility
Declined offer
Yes
No
Accepted offer
Declined offer

inicio

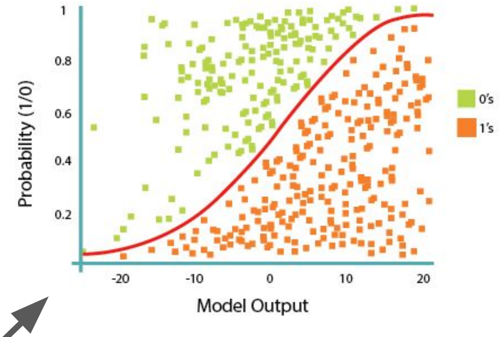
Quiero explicar las decisiones del modelo?

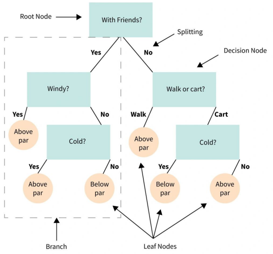
si

Es el modelo interpretable por su diseño?

si

KNN , LR , DECISION TREE LINEAR





no

¿Necesita de otro modelo para interpretarlo?

inicio

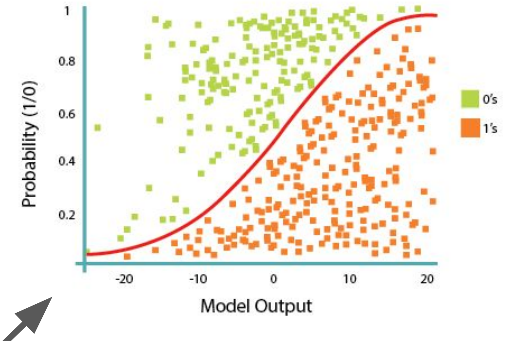
Quiero explicar las decisiones del modelo?

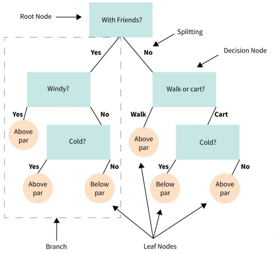
si

Es el modelo interpretable por su diseño?

si

KNN , LR , DECISION TREE LINEAR





no

¿Necesita de otro modelo para interpretarlo?

si

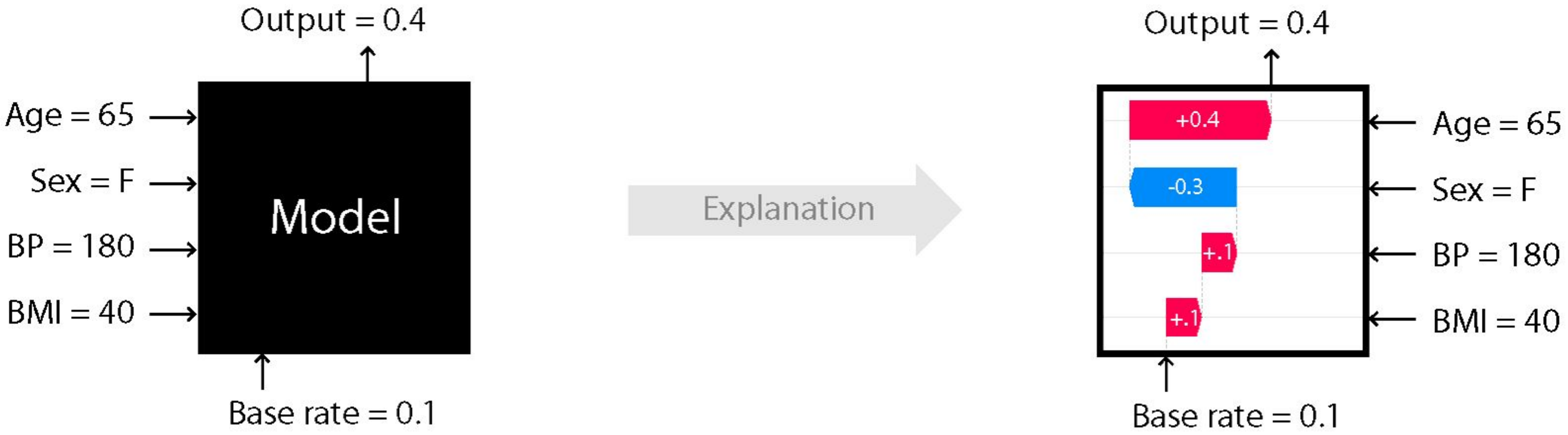
Model Agnostic Explanations

Example based (Counterfactual, Adversarial, Influence)

Model-agnostic explanations

Teoría de juegos:
Como un cambio en los valores de los features cambia el output promedio marginal.

SHAP



LIME

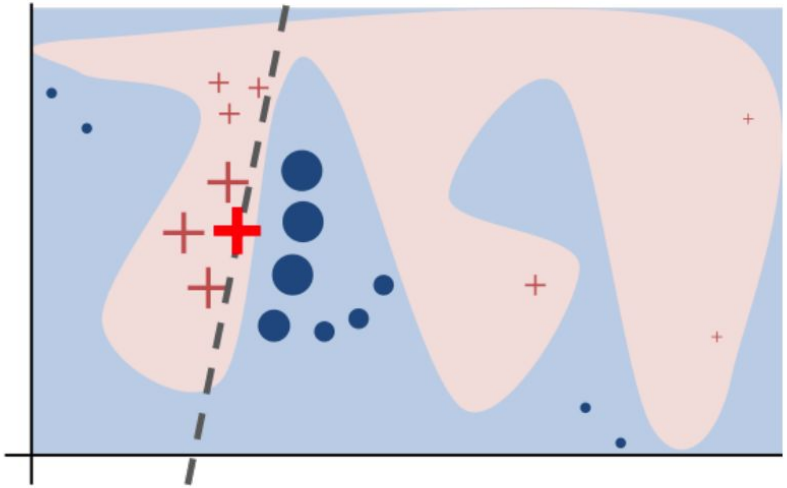
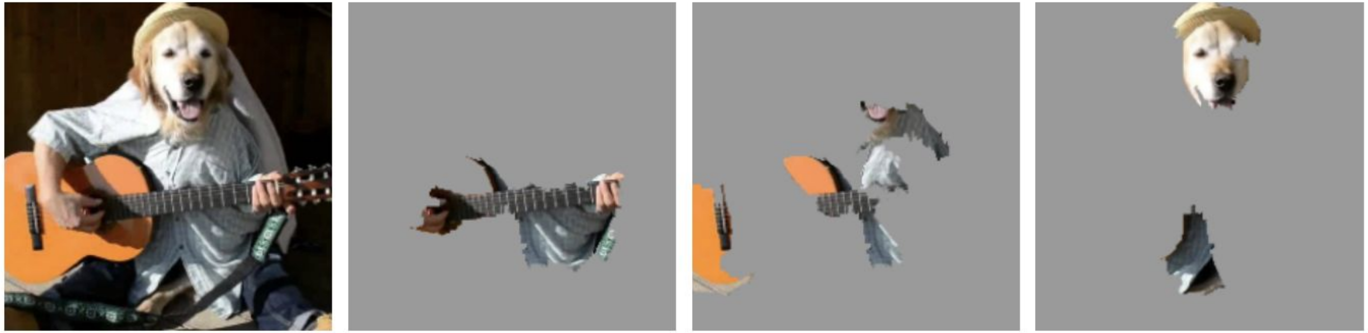


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

LIME



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Modelo predijo 3 clases:

- Guitarra electrica (**p=**0.32) → pixeles del mástil de la guitarra
- Guitarra acustica (**p=**0.24) → pixeles del cuerpo de la guitarra
- Labrador (**p=**0.21) → cara del perro labrador.

Example-based explanations

Counterfactual

¿Qué tendría que cambiar para obtener un resultado diferente?

Ejemplo:

Si tu ingreso anual hubiera sido $X más alto, tu solicitud de crédito habría sido aprobada.

Así podemos encontrar si hay sesgos del modelo por ciertas características.

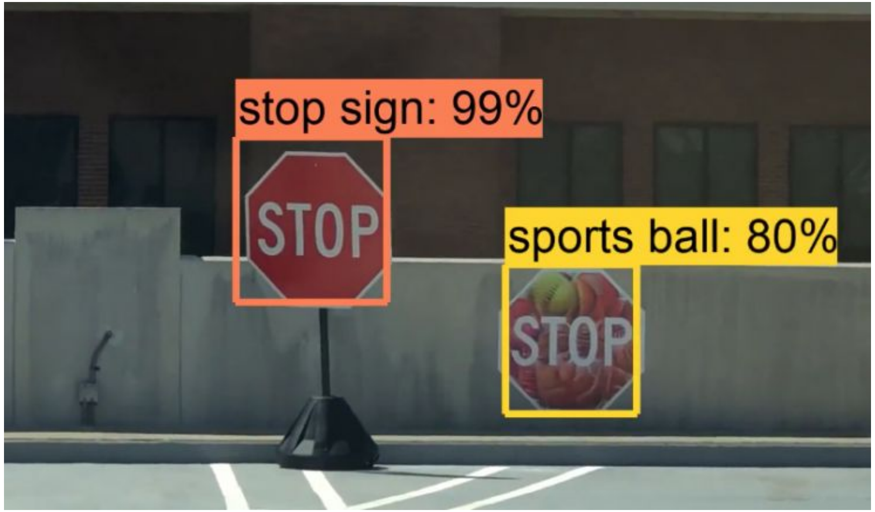
stop sign
Confidence: 0.9153

Adversarial perturbation

flowerpot
Confidence: 0.8374

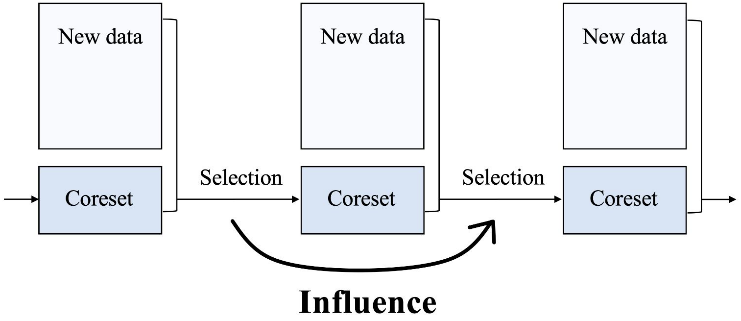
Ejemplos adversarios:

entradas perturbadas imperceptibles que induzcan a errores en un modelo



Funciones de influencia (Influence functions)

Las funciones de influencia se utilizan para analizar la sensibilidad de un estimador (por ejemplo, la media o la mediana) ante perturbaciones en el conjunto de datos.



inicio

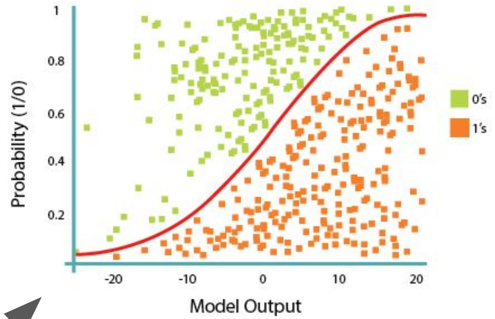
Quiero explicar las decisiones del modelo?

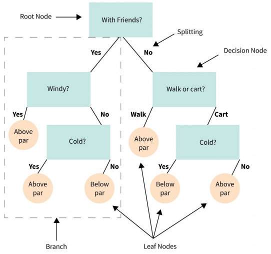
si

Es el modelo interpretable por su diseño?

si

KNN , LR , DECISION TREE LINEAR





no

¿Necesita de otro modelo para interpretarlo?

si

Model Agnostic Explanations

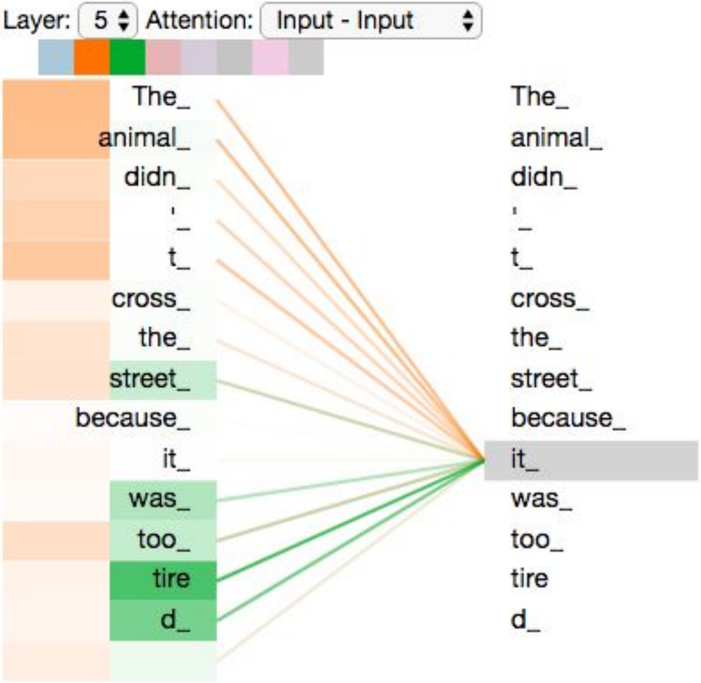
SHAP

no

- attention
- gradient saliency
- integrated gradients

Attention

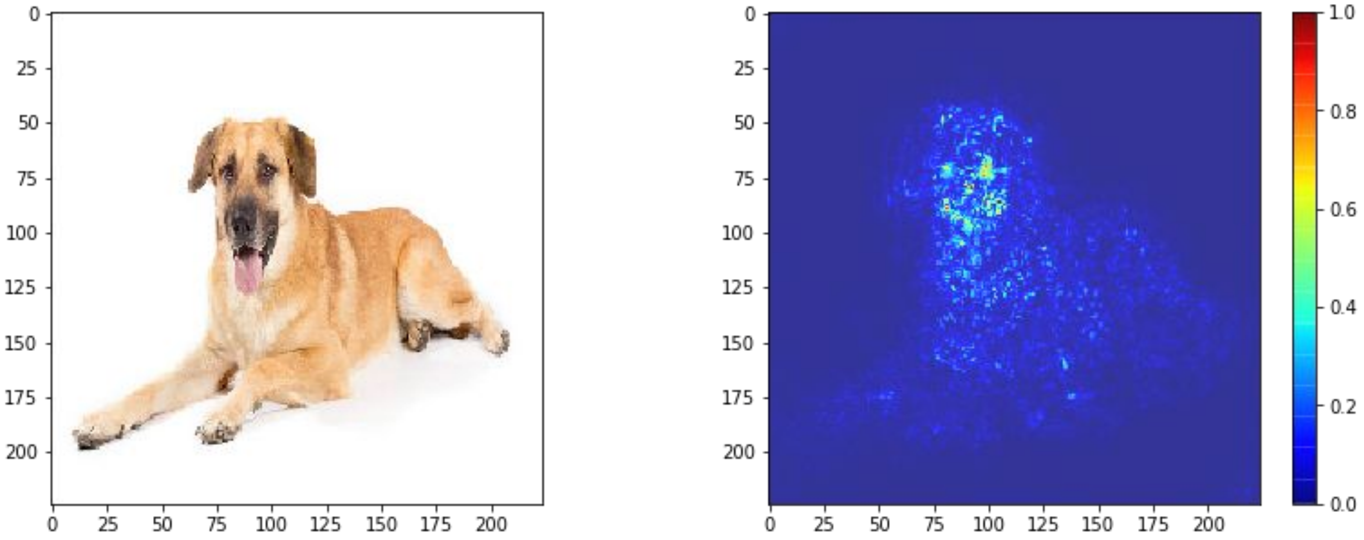
Estudiar la atencion del modelo al generar inferencia sobre nuevos ejemplos.



Layer: 5 Attention: Input - Input

Gradient saliency maps

Estudiar cuáles partes de la entrada (ej. imagen) tienen mayor influencia en el gradiente de la red neuronal.

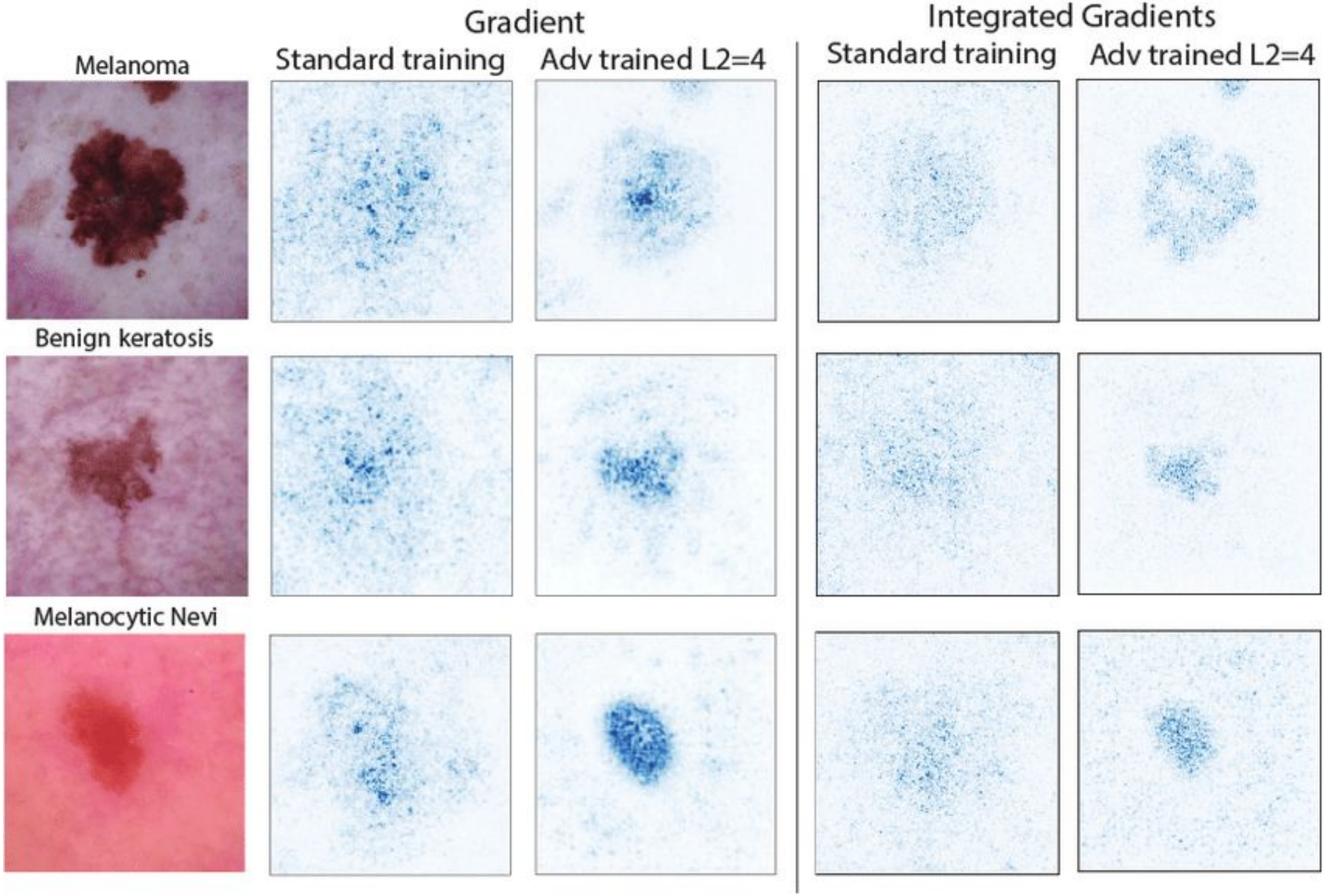


Integrated gradients

Se elige un punto de referencia neutro, como una imagen en blanco o negro.

Luego, se traza un camino desde ese punto hasta la entrada real.

A lo largo de este camino, se suman los gradientes del modelo para determinar la importancia de cada característica.


Gradient
Integrated Gradients
Standard training
Adv trained L2=4
Standard training
Adv trained L2=4
Melanoma
Benign keratosis
Melanocytic Nevi

inicio

Quiero explicar las decisiones del modelo?

no

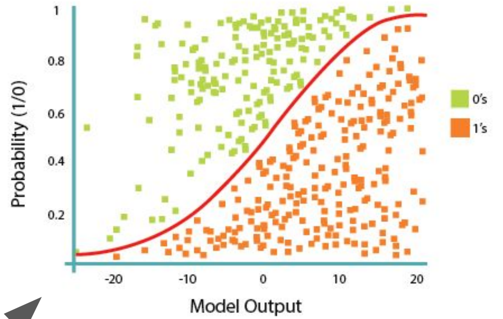
entender cómo aprende el modelo
ej. feature visualization.

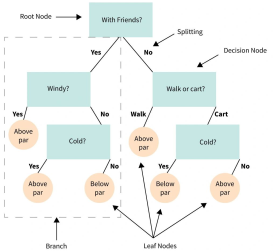
si

Es el modelo interpretable por su diseño?

si

KNN , LR ,
DECISION
TREE
LINEAR





no

¿Necesita de otro modelo para interpretarlo?

no

- attention
- gradient saliency (feat importance)
- integrated gradients

si

Model Agnostic Explanations

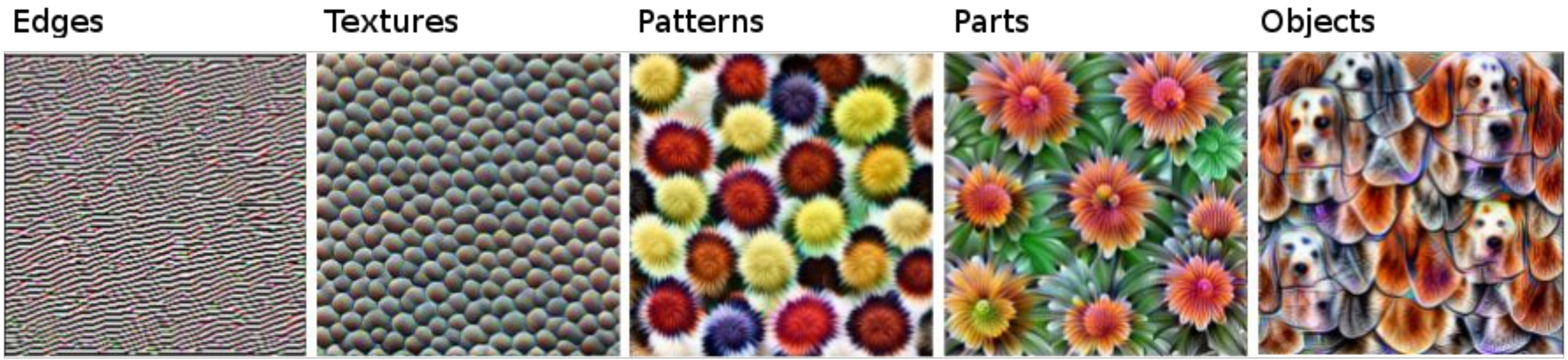
SHAP

si

Example-based

counterfactual
adversarial

Feature visualization



Explorar que aprende cada una de las neuronas.

fuente: https://distill.pub/2017/feature-visualization/

Explainable Recommendation

La recomendación explicable busca desarrollar modelos que no solo generen recomendaciones de alta calidad, sino también explicaciones intuitivas.

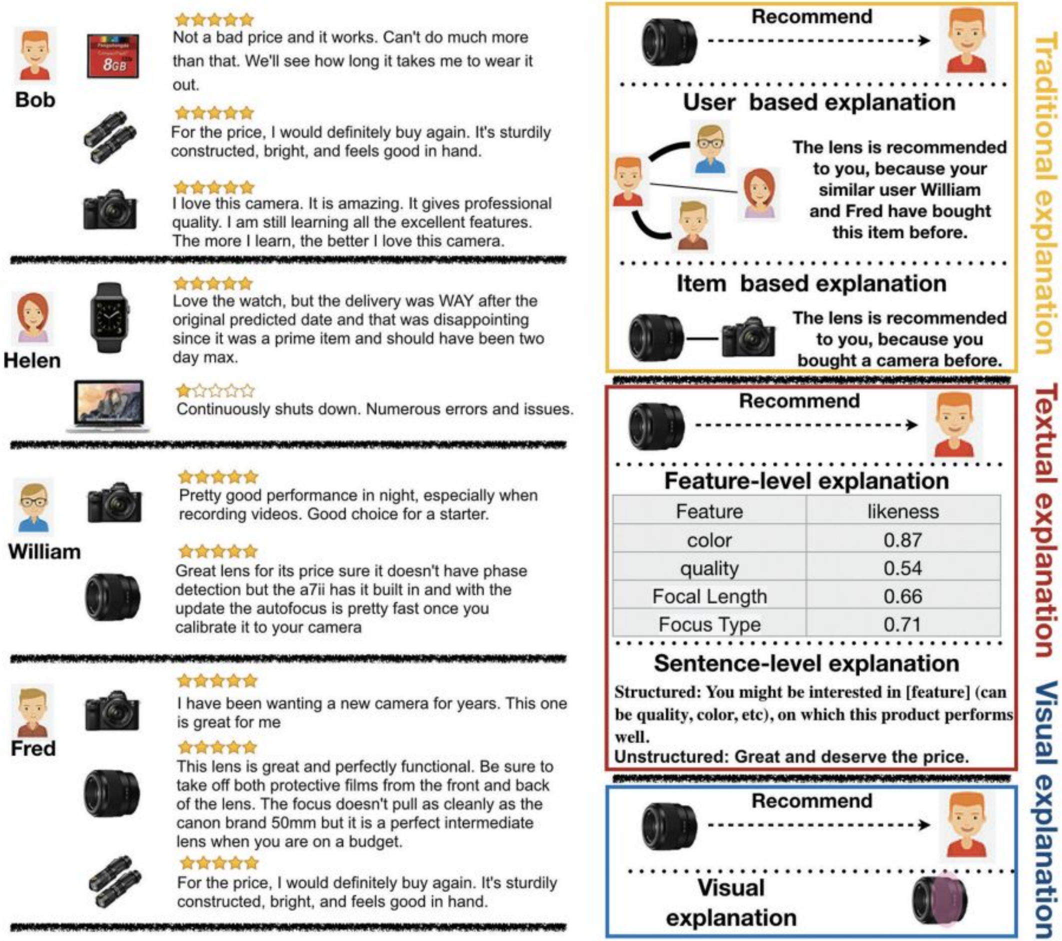
Las explicaciones pueden ser

- post-hoc
- provenir directamente de un modelo interpretable o transparente

Aborda el problema del por qué, ayudando a los humanos (usuarios o diseñadores de sistemas) a entender por qué ciertos ítems son recomendados por el algoritmo.

Beneficios

- Mejora la transparencia de los sistemas de recomendación.
- Aumenta la persuasión y efectividad de las recomendaciones.
- Fomenta la confianza y satisfacción del usuario.
- Facilita a los diseñadores del sistema una mejor depuración.



★★★★★
Not a bad price and it works. Can't do much more than that. We'll see how long it takes me to wear it out.

★★★★★
For the price, I would definitely buy again. It's sturdily constructed, bright, and feels good in hand.

★★★★★
I love this camera. It is amazing. It gives professional quality. I am still learning all the excellent features. The more I learn, the better I love this camera.

★★★★★
Love the watch, but the delivery was WAY after the original predicted date and that was disappointing since it was a prime item and should have been two day max.

★☆☆☆☆
Continuously shuts down. Numerous errors and issues.

★★★★★
Pretty good performance in night, especially when recording videos. Good choice for a starter.

★★★★★
Great lens for its price sure it doesn't have phase detection but the a7ii has it built in and with the update the autofocus is pretty fast once you calibrate it to your camera

★★★★★
I have been wanting a new camera for years. This one is great for me

★★★★★
This lens is great and perfectly functional. Be sure to take off both protective films from the front and back of the lens. The focus doesn't pull as cleanly as the canon brand 50mm but it is a perfect intermediate lens when you are on a budget.

★★★★★
For the price, I would definitely buy again. It's sturdily constructed, bright, and feels good in hand.

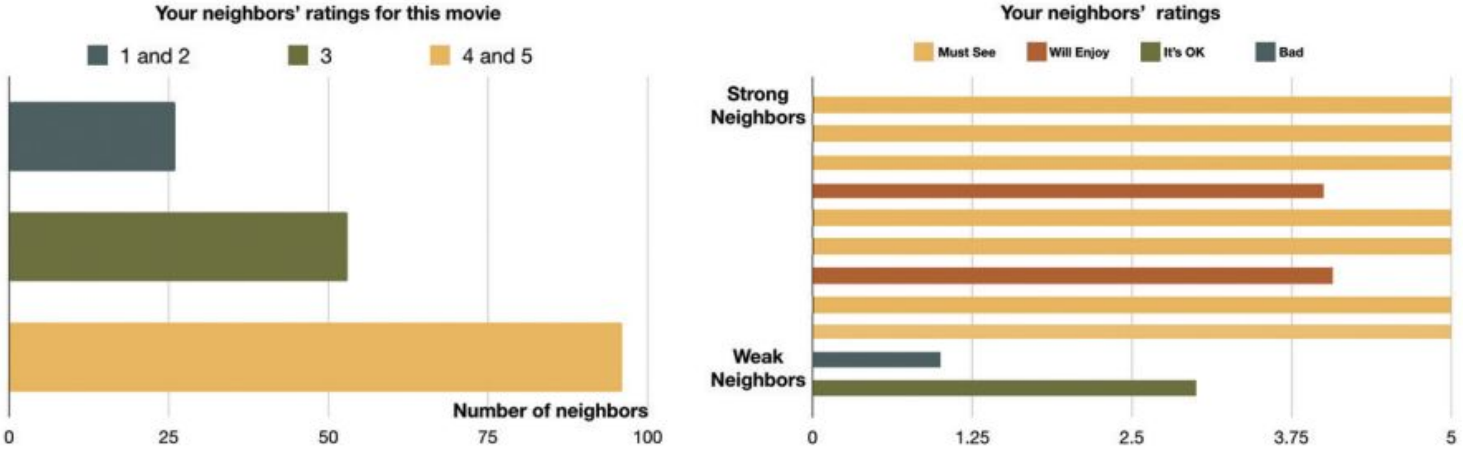
The lens is recommended to you, because your similar user William and Fred have bought this item before.

The lens is recommended to you, because you bought a camera before.

Feature	likeness
color	0.87
quality	0.54
Focal Length	0.66
Focus Type	0.71

Structured: You might be interested in [feature] (can be quality, color, etc), on which this product performs well.

Unstructured: Great and deserve the price.


Your neighbors' ratings for this movie
1 and 2
3
4 and 5
Number of neighbors
0
25
50
75
100
Your neighbors' ratings
Must See
Will Enjoy
It's OK
Bad
Strong Neighbors
Weak Neighbors
0
1.25
2.5
3.75
5

(a)

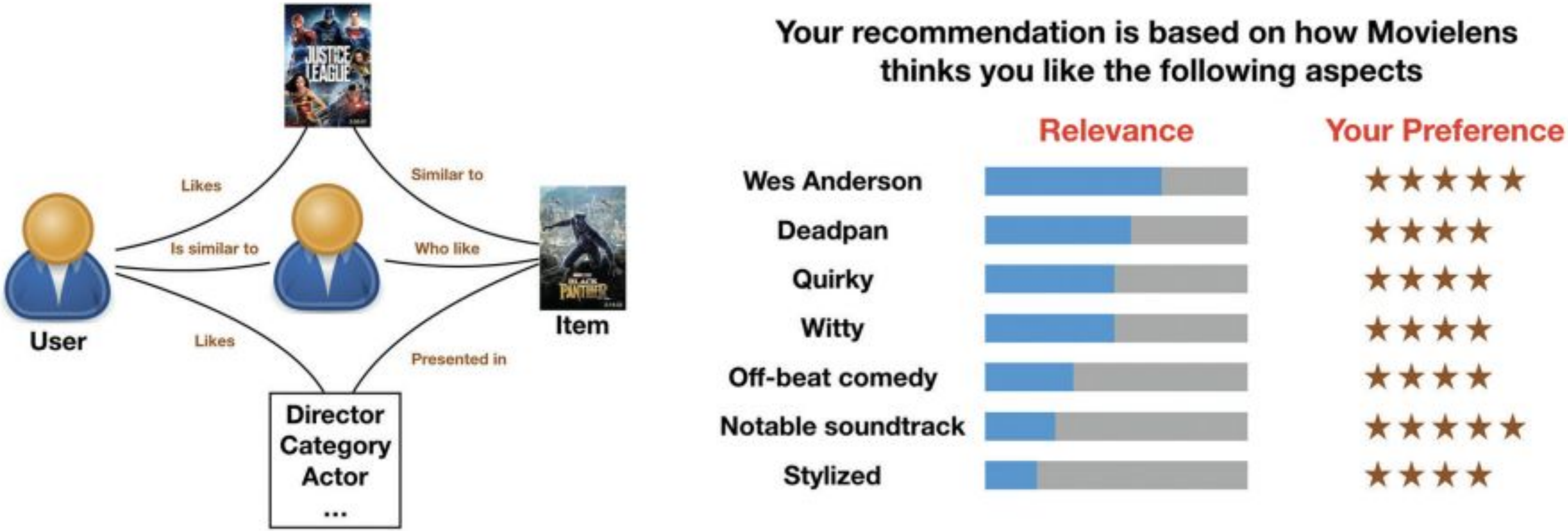
(b)

Your rating for similar movies

★★★★★ ★★★★ ★★★★★

Your neighbors' rating for this movie

Rating	Number of Neighbors
★★★★★	2
★★★★	3
★★★	4
★★	0
★	0

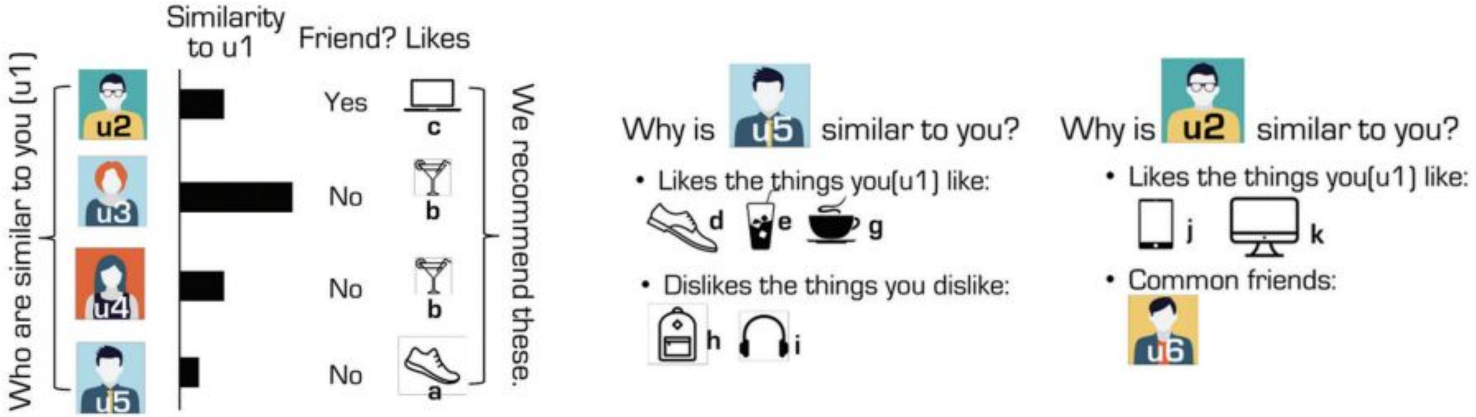

Likes
Similar to
Is similar to
Who like
User
Item
Likes
Presented in
Director
Category
Actor
...
Your recommendation is based on how Movielens thinks you like the following aspects
Relevance
Your Preference
Wes Anderson ★★★★★
Deadpan ★★★★
Quirky ★★★★
Witty ★★★★
Off-beat comedy ★★★★
Notable soundtrack ★★★★★
Stylized ★★★★

(a) (b)

Amazon Books

Recommended

This is well written book with a very good detail of a person that love his dog but didn't restrain his freedom. I think that I relive the joy of my experiences with my dogs, a Labrador and a Siberian Husky. Both were rescued, one from the shelter and the other from the street. After 4 months with me, the owner of the husky appeared and I returned the dog. Two weeks later the dog escaped and returned to my house. He decided who will be his owner. The author described with details the relationship of them, concerns, disappointments and health issues. The final chapter was a surprise that I am still enjoying.



Why is u5 similar to you?

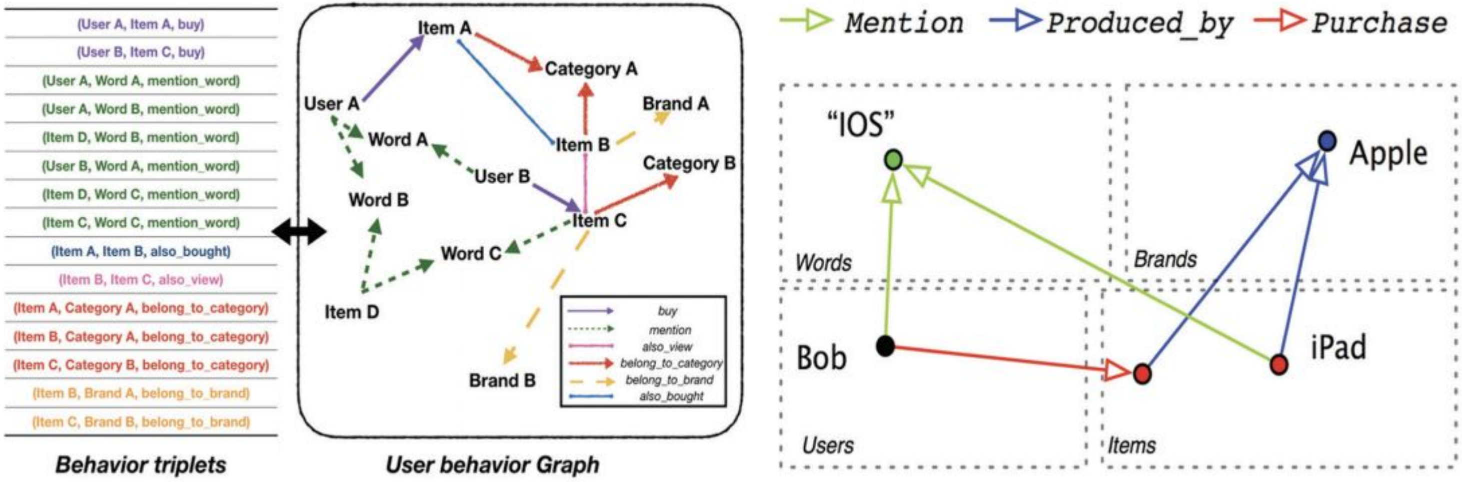
- Likes the things you(u1) like:

- Dislikes the things you dislike:

Why is u2 similar to you?

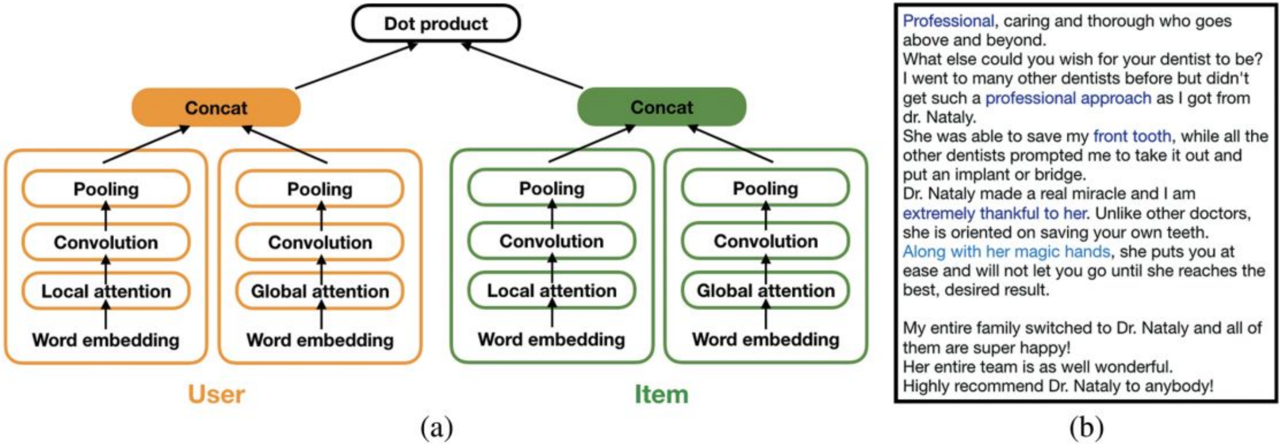
- Likes the things you(u1) like:

- Common friends:


(User A, Item A, buy)
(User B, Item C, buy)
(User A, Word A, mention_word)
(User A, Word B, mention_word)
(Item D, Word B, mention_word)
(User B, Word A, mention_word)
(Item D, Word C, mention_word)
(Item C, Word C, mention_word)
(Item A, Item B, also_bought)
(Item B, Item C, also_view)
(Item A, Category A, belong_to_category)
(Item B, Category A, belong_to_category)
(Item C, Category B, belong_to_category)
(Item B, Brand A, belong_to_brand)
(Item C, Brand B, belong_to_brand)
Behavior triplets
Item A
Category A
User A
Brand A
Word A
Item B
User B
Category B
Word B
Item C
Word C
Item D
Brand B
buy
mention
also_view
belong_to_category
belong_to_brand
also_bought
User behavior Graph
Mention
Produced_by
Purchase
"IOS"
Apple
Words
Brands
Bob
iPad
Users
Items

(a) A knowledge graph of users and items

(b) Extracting explanation path



(a)

Professional, caring and thorough who goes above and beyond.
What else could you wish for your dentist to be?
I went to many other dentists before but didn't get such a professional approach as I got from dr. Nataly.
She was able to save my front tooth, while all the other dentists prompted me to take it out and put an implant or bridge.
Dr. Nataly made a real miracle and I am extremely thankful to her. Unlike other doctors, she is oriented on saving your own teeth.
Along with her magic hands, she puts you at ease and will not let you go until she reaches the best, desired result.

My entire family switched to Dr. Nataly and all of them are super happy!
Her entire team is as well wonderful.
Highly recommend Dr. Nataly to anybody!

(b)

Figure 3.8: (a) The dual-attention architecture to extract user and item representations. A user document and an item document are fed into the user network (left) and item network (right). (b) The model generates attention scores for each review and highlights the high attention words as explanations (Seo *et al.*, 2017).