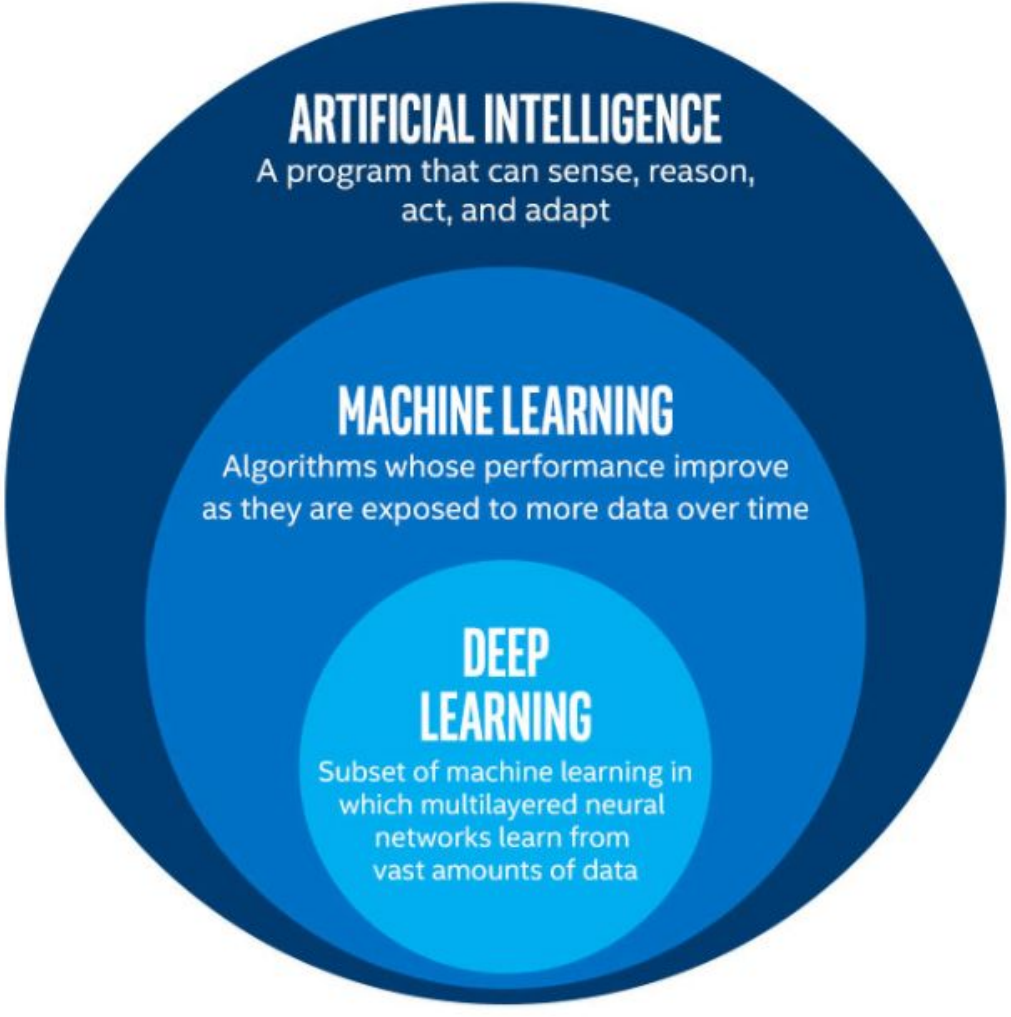
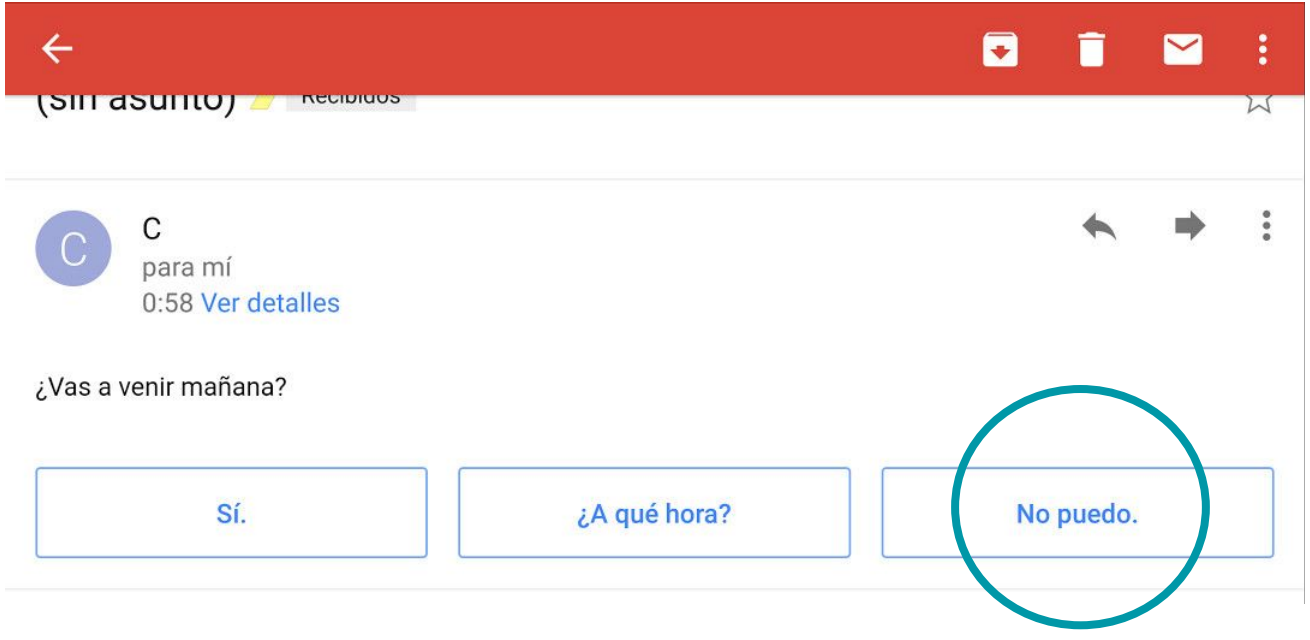
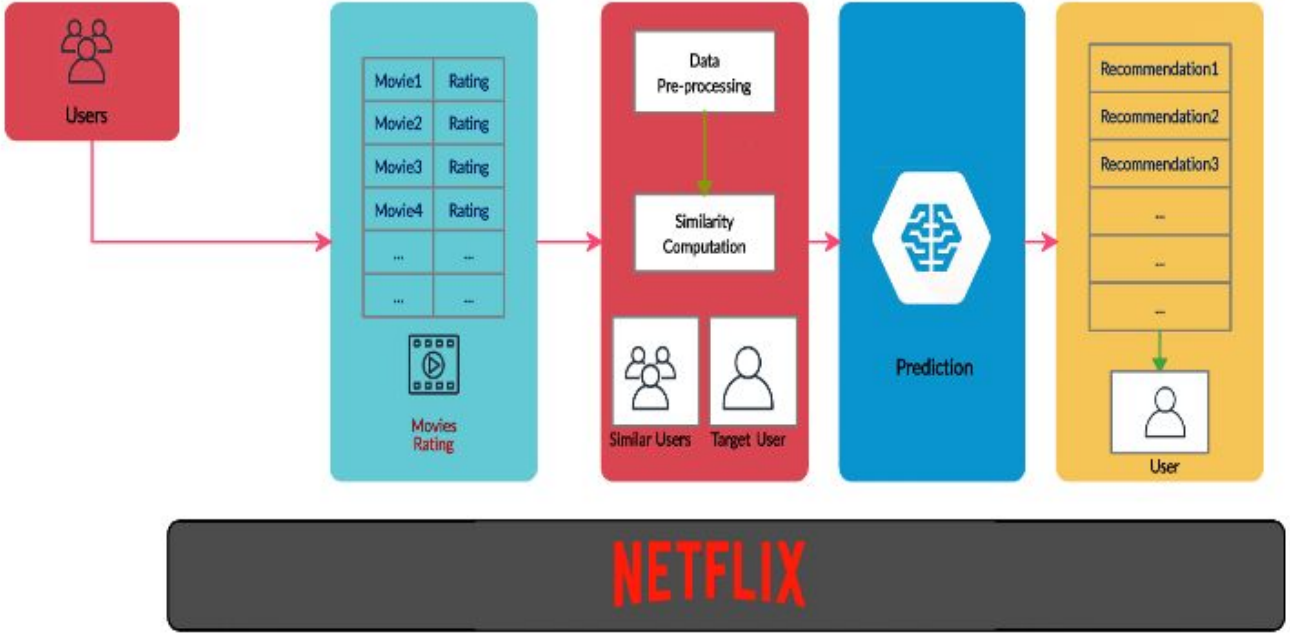
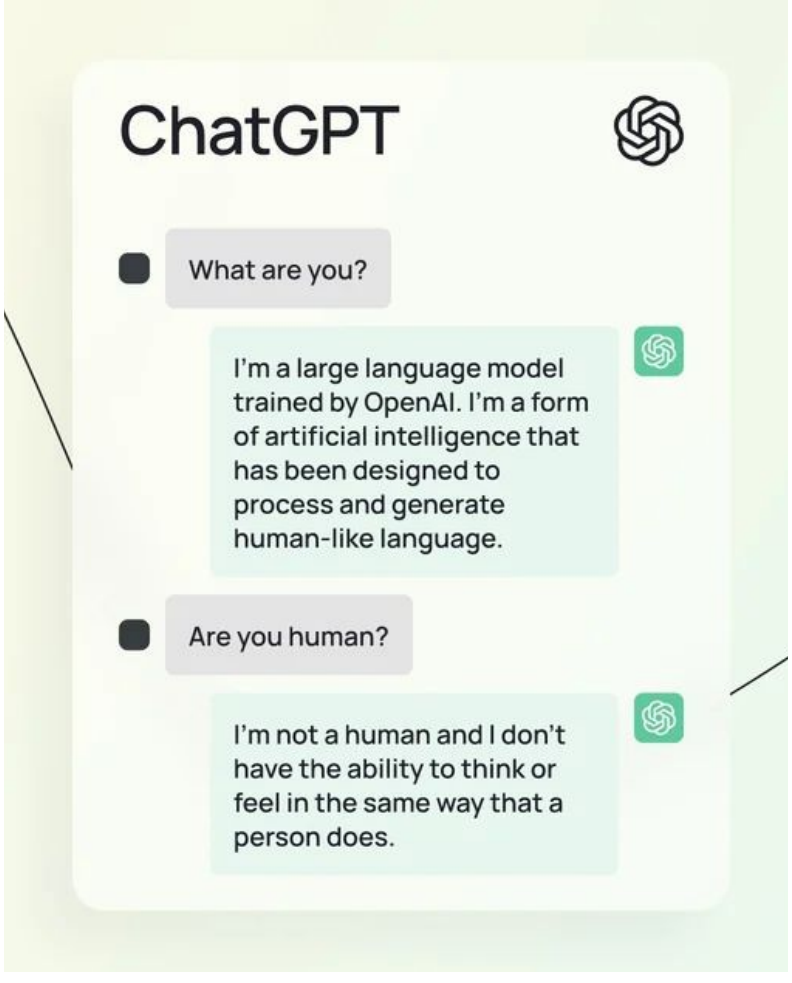
Revelando los secretos de la caja negra: Una inmersión en las profundidades de la inteligencia artificial


ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act, and adapt
MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time
DEEP LEARNING
Subset of machine learning in which multilayered neural networks learn from vast amounts of data

IA en nuestro cotidiano


(sin asunto)
Recibidos
C
C
para mí
0:58 Ver detalles
¿Vas a venir mañana?
Sí.
¿A qué hora?
No puedo.


Users
Movie1
Rating
Movie2
Rating
Movie3
Rating
Movie4
Rating
...
...
...
...
Movies
Rating
Data
Pre-processing
Similarity
Computation
Similar Users
Target User
Prediction
Recommendation1
Recommendation2
Recommendation3
...
...
...
User
NETFLIX


ChatGPT
What are you?
I'm a large language model trained by OpenAI. I'm a form of artificial intelligence that has been designed to process and generate human-like language.
Are you human?
I'm not a human and I don't have the ability to think or feel in the same way that a person does.

We have removed your post from Facebook

Post unavailable

We are unable to show content that goes against our Community Standards on hate speech.

To protect others' privacy and keep everyone safe from harmful content, we do not show content that violates Community Standards on hate speech.

What happened

Your appeal was reviewed and your post does not follow our Community Standards for hate speech.

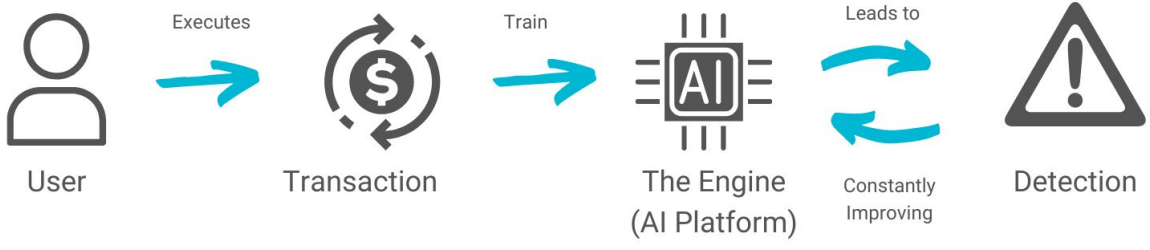
What you can do

You can go to Account status to see how content violations can affect you.

Close

TRADITIONAL RULE-BASED APPROACH

MACHINE LEARNING APPROACH



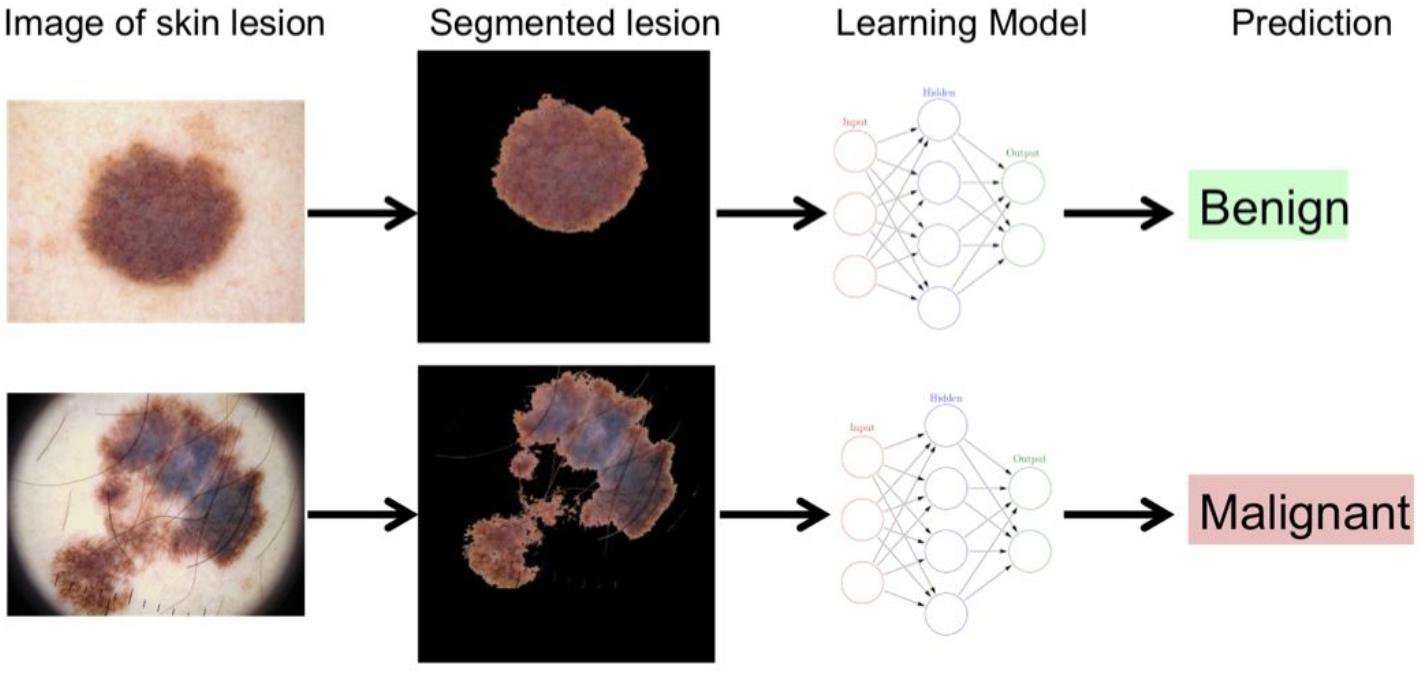
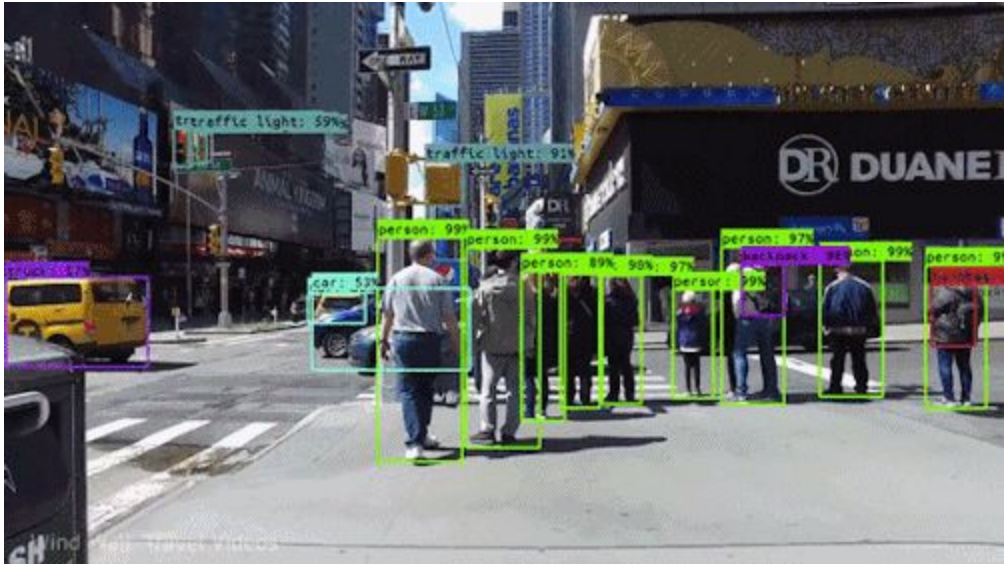
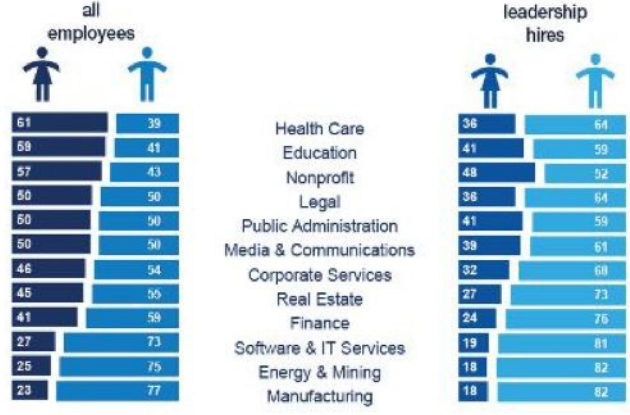


Image of skin lesion
Segmented lesion
Learning Model
Prediction
Input
Hidden
Output
Benign
Malignant

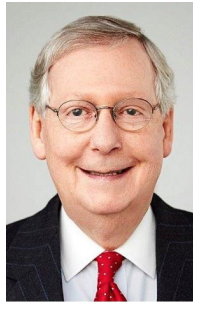

treaffic light: 59%
traffic light: 91%
DUANE
person: 99%
person: 99%
person: 89%
car: 53%
person: 97%
person: 99%

Limitaciones y riesgos de la IA



Source: LinkedIn data featured in the
Global Gender Gap Report 2017, World Economic Forum


Tony "Abolish ICE" Arcieri
@bascule
Trying a horrible experiment...
Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?
8:05 AM · Sep 20, 2020
193.5K
2.8K
Share this Tweet



Explore Images Related to Crime Generated by Stable Diffusion

A color photograph of a **drug dealer**

Explore Images Related to Crime Generated by Stable Diffusion

A color photograph of an **inmate**

Explore Images Related to Crime Generated by Stable Diffusion

A color photograph of a **terrorist**

Como evitar esto?

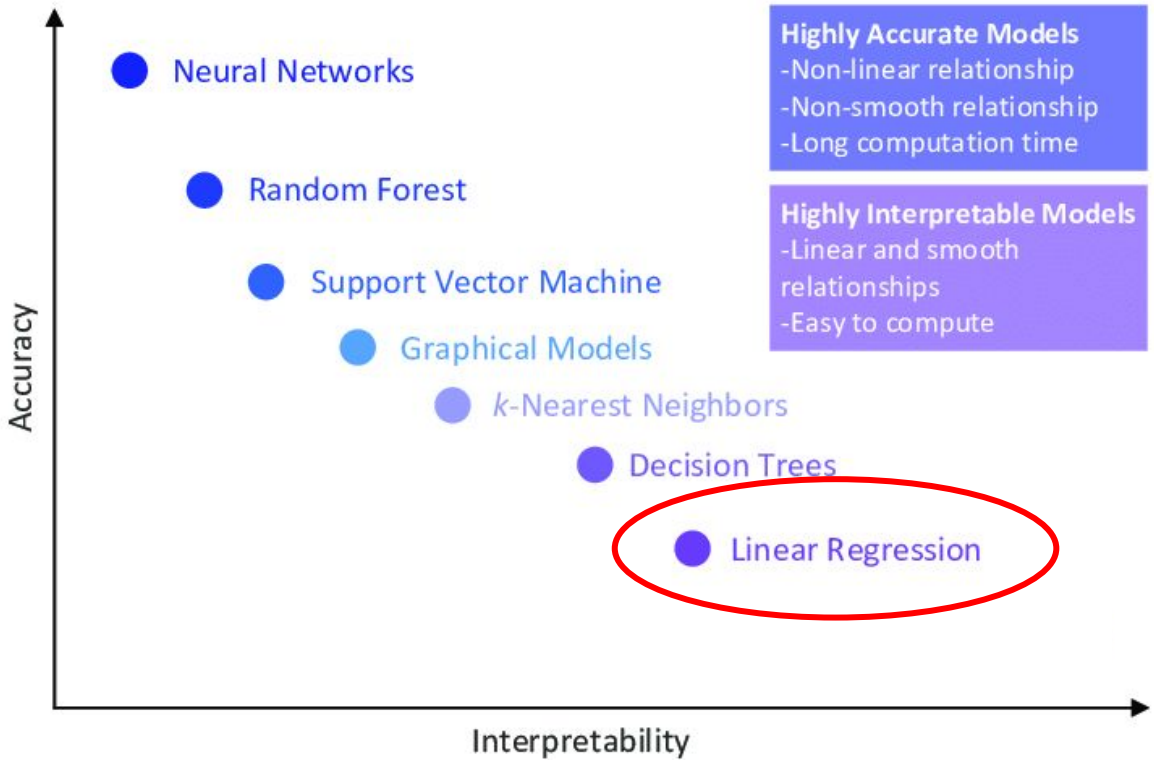
Interpretabilidad de Modelos

Modelos Paramétricos

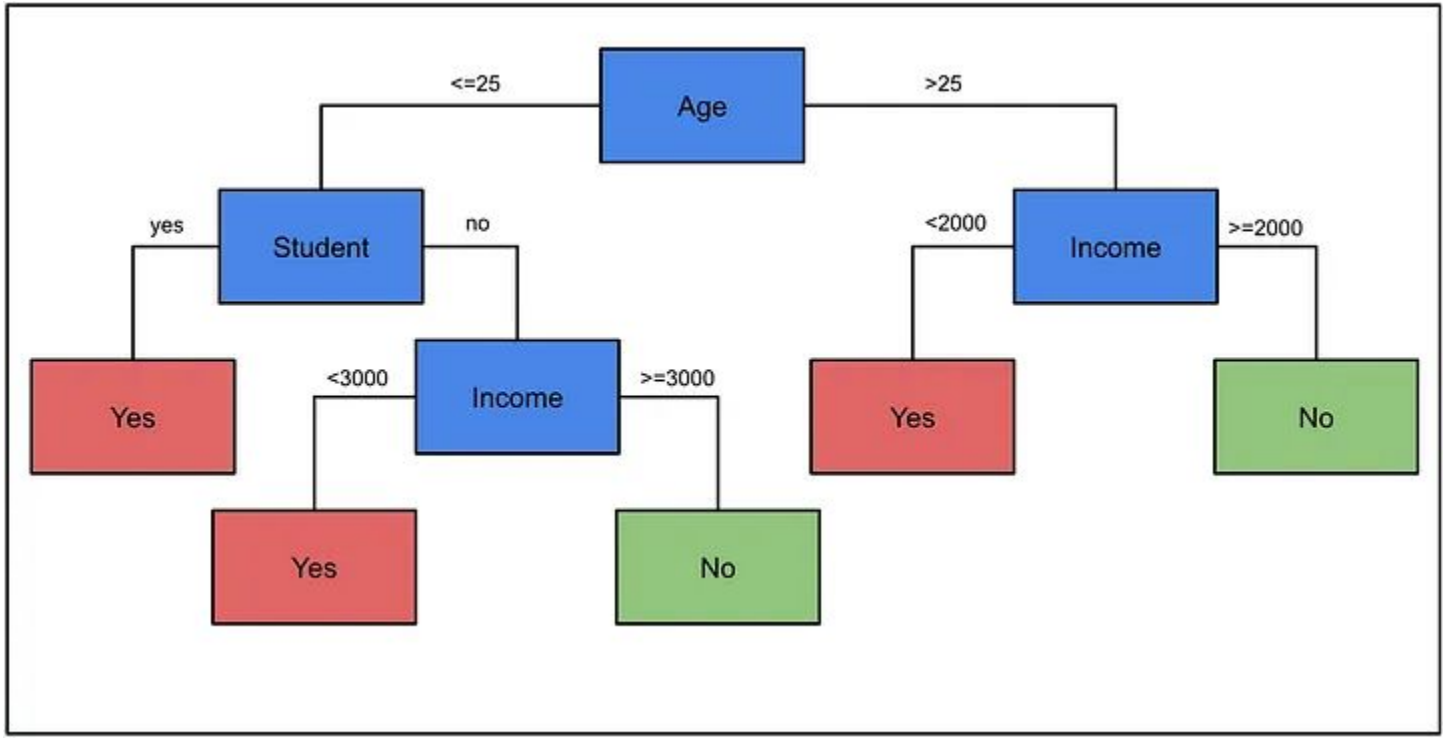
- La Regresión Lineal es un modelo paramétrico, lo que significa que la hipótesis se describe en términos de coeficientes que ajustamos para mejorar la precisión del modelo.

$$\mathbf{Y} = 100*\mathbf{age} + 10*\mathbf{income} + 200$$

$$Y = X_0 + B_1X_1 + B_2X_2 + \cdots + B_nX_n$$

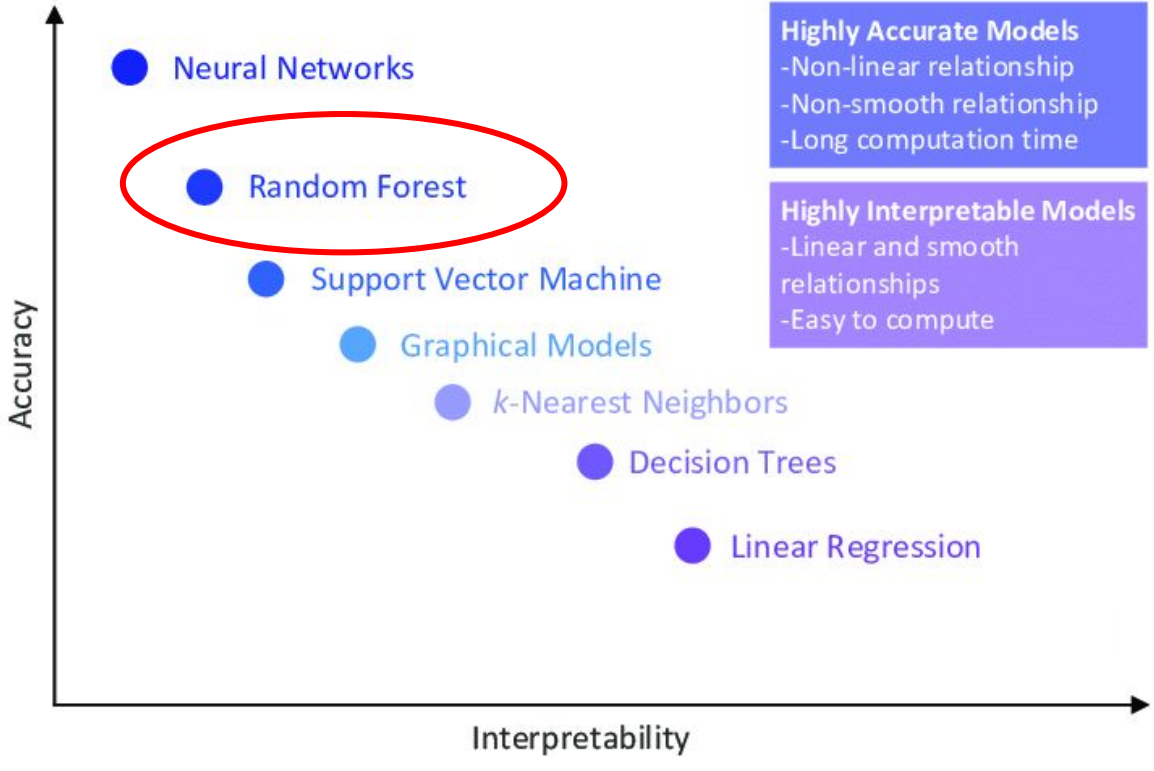

Neural Networks
Random Forest
Support Vector Machine
Graphical Models
k-Nearest Neighbors
Decision Trees
Linear Regression
Accuracy
Interpretability
Highly Accurate Models
-Non-linear relationship
-Non-smooth relationship
-Long computation time
Highly Interpretable Models
-Linear and smooth relationships
-Easy to compute

Modelos No Paramétricos



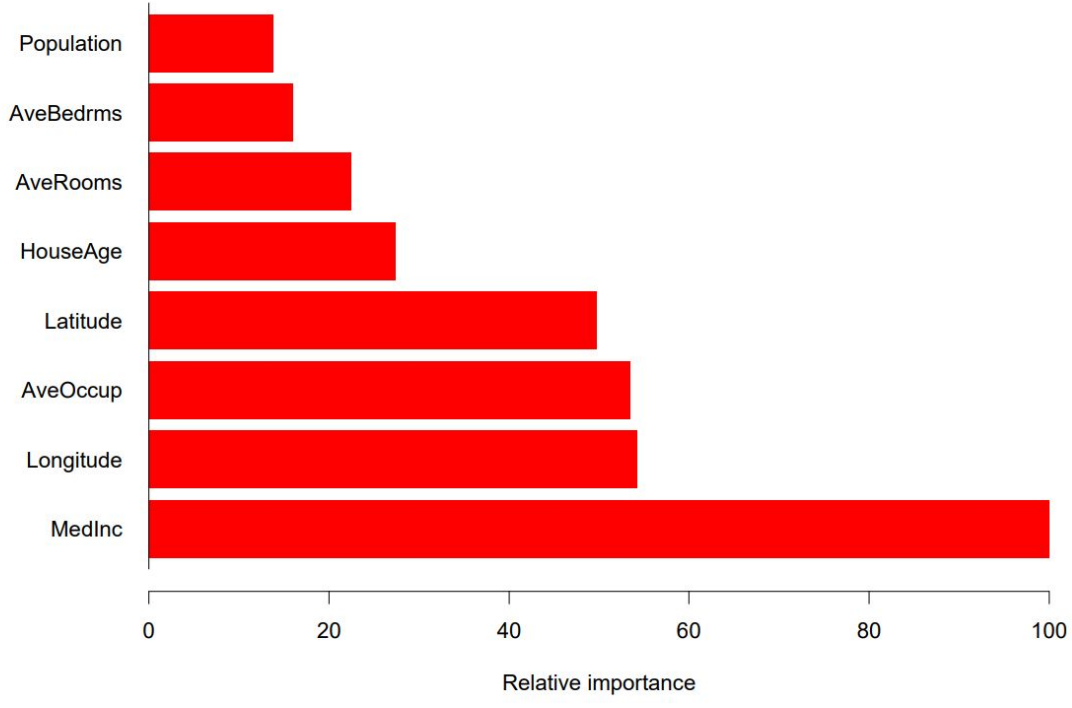
- Modelos como **KNN** o **Naive Bayes** no tienen parámetros que permitan entender la importancia de las variables ni su interpretación.

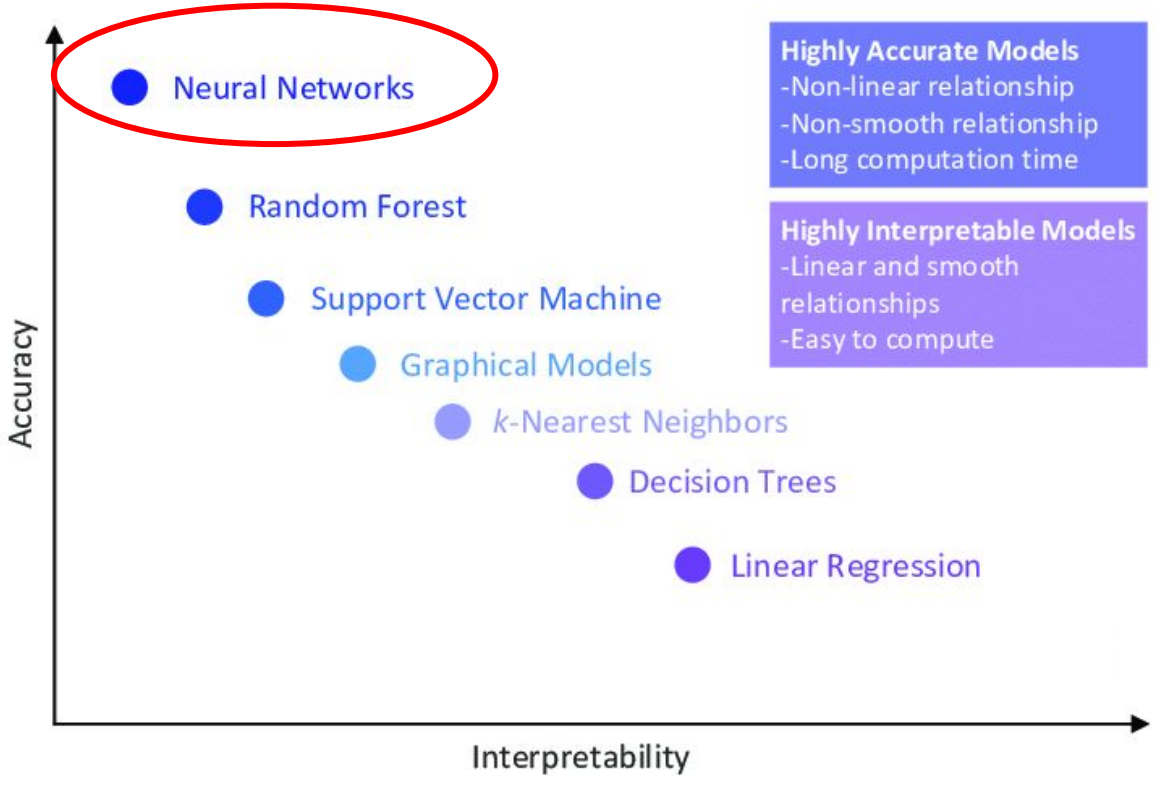
- Los modelos basados en árboles (**Random Forest, XGBOOST**) son otro ejemplo, no tenemos coeficientes. Sin embargo aún nos puede interesar saber cuáles features son más importantes.

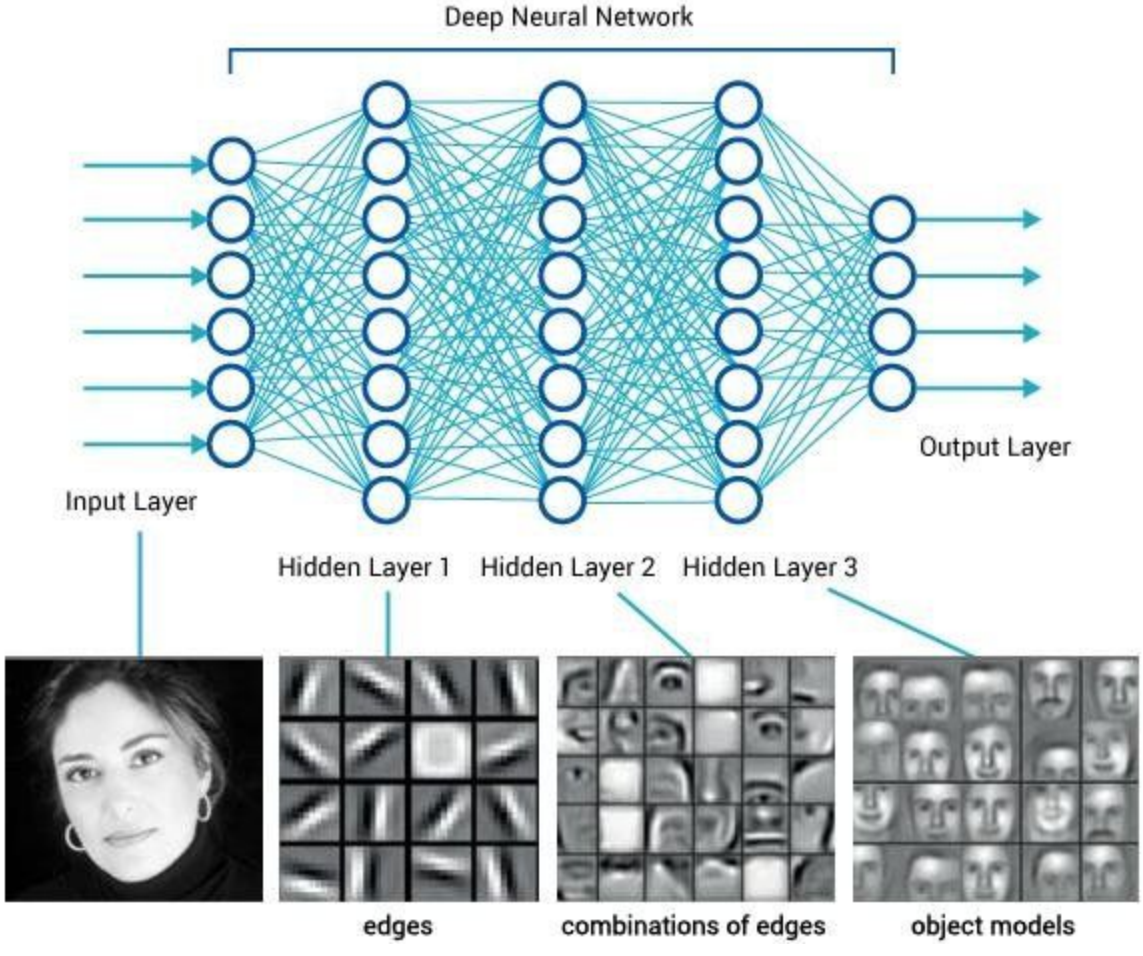

Neural Networks
Random Forest
Support Vector Machine
Graphical Models
k-Nearest Neighbors
Decision Trees
Linear Regression
Highly Accurate Models
-Non-linear relationship
-Non-smooth relationship
-Long computation time
Highly Interpretable Models
-Linear and smooth relationships
-Easy to compute
Accuracy
Interpretability

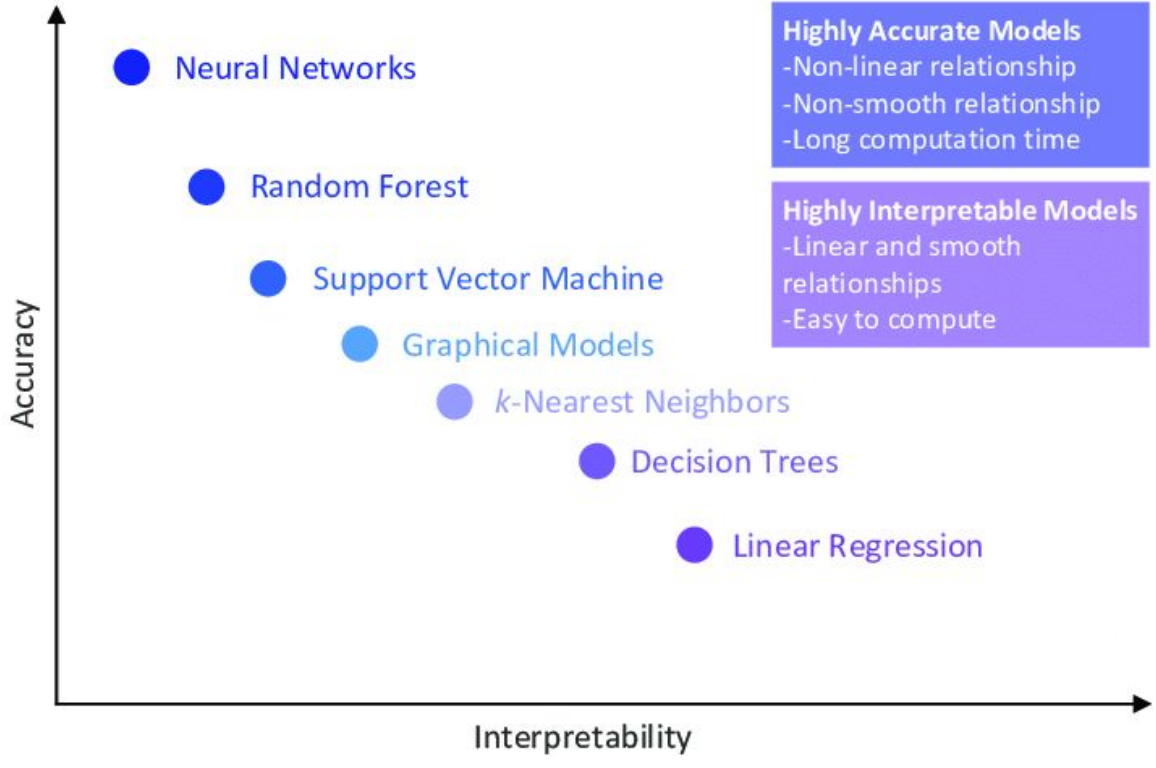
- Un random forest es un conjunto de árboles entrenados en muestras aleatorias y subconjuntos aleatorios de features.

- Podemos **calcular la importancia de los features de cada árbol y luego promediar las importancias en todo el ensamble.**


Population
AveBedrms
AveRooms
HouseAge
Latitude
AveOccup
Longitude
MedInc
0
20
40
60
80
100
Relative importance


Neural Networks
Random Forest
Support Vector Machine
Graphical Models
k-Nearest Neighbors
Decision Trees
Linear Regression
Highly Accurate Models
-Non-linear relationship
-Non-smooth relationship
-Long computation time
Highly Interpretable Models
-Linear and smooth relationships
-Easy to compute
Accuracy
Interpretability


Deep Neural Network
Input Layer
Hidden Layer 1
Hidden Layer 2
Hidden Layer 3
Output Layer
edges
combinations of edges
object models


Neural Networks
Random Forest
Support Vector Machine
Graphical Models
k-Nearest Neighbors
Decision Trees
Linear Regression
Highly Accurate Models
-Non-linear relationship
-Non-smooth relationship
-Long computation time
Highly Interpretable Models
-Linear and smooth relationships
-Easy to compute
Accuracy
Interpretability

Data

Results

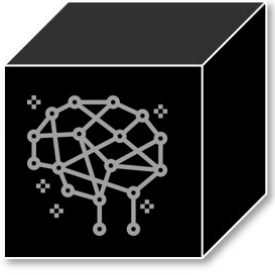
This is an insect!

Data

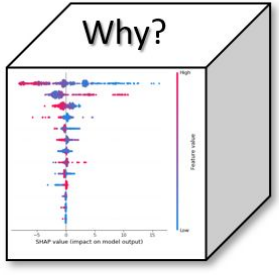
Results

This is an insect!

Data



Explanation



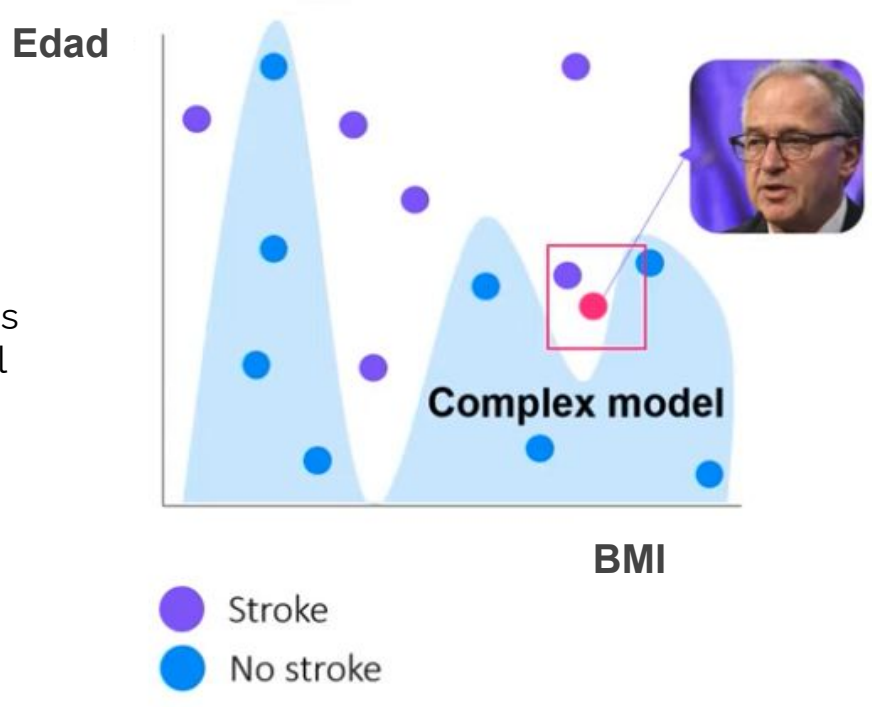
Because it has 6 legs

Results

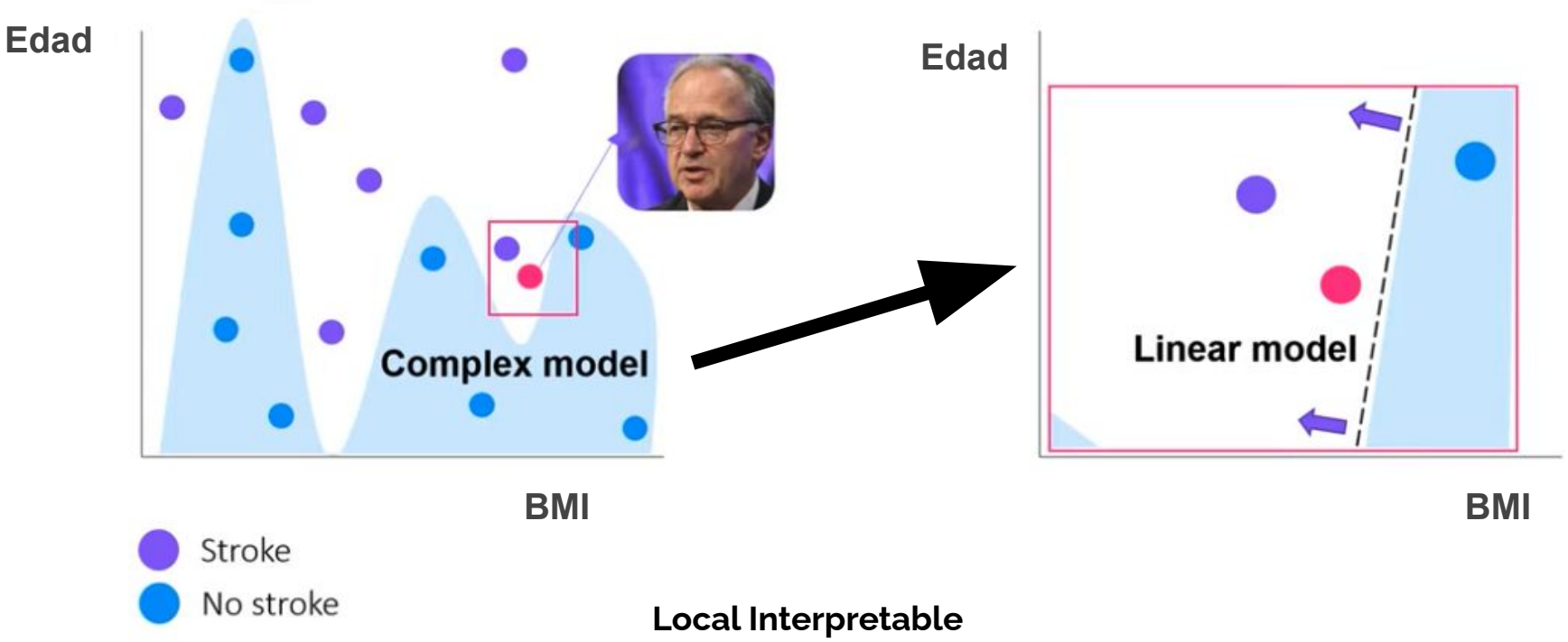
This is an insect!

Explicabilidad Local con LIME

Complejidad: Nuestro principal problema es la alta no linealidad de nuestros modelos, como no tenemos parámetros, es muy difícil darle una interpretación general al modelo.

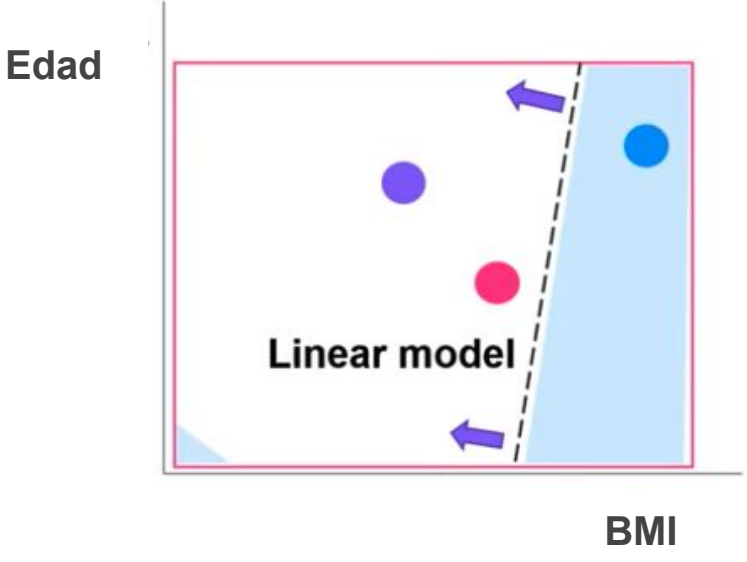


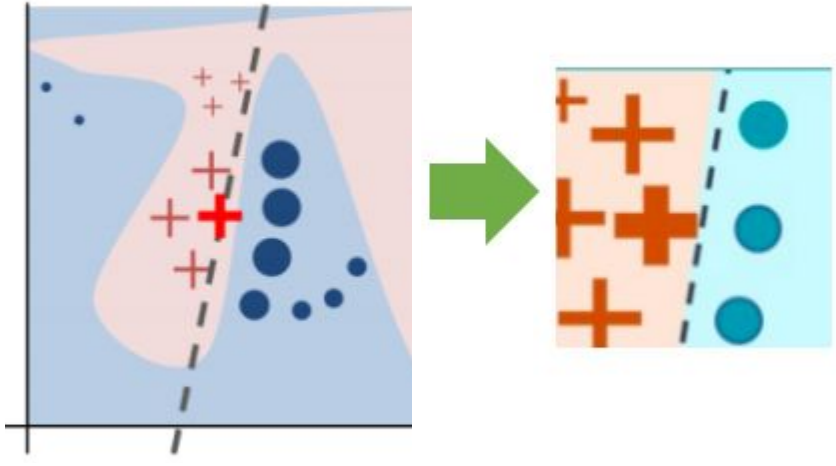
https://github.com/marcotcr/lime



Local Interpretable Model-Agnostic Explanations (LIME)

Localmente: Alrededor de la predicción que buscamos explicar podemos recuperar un comportamiento lineal.



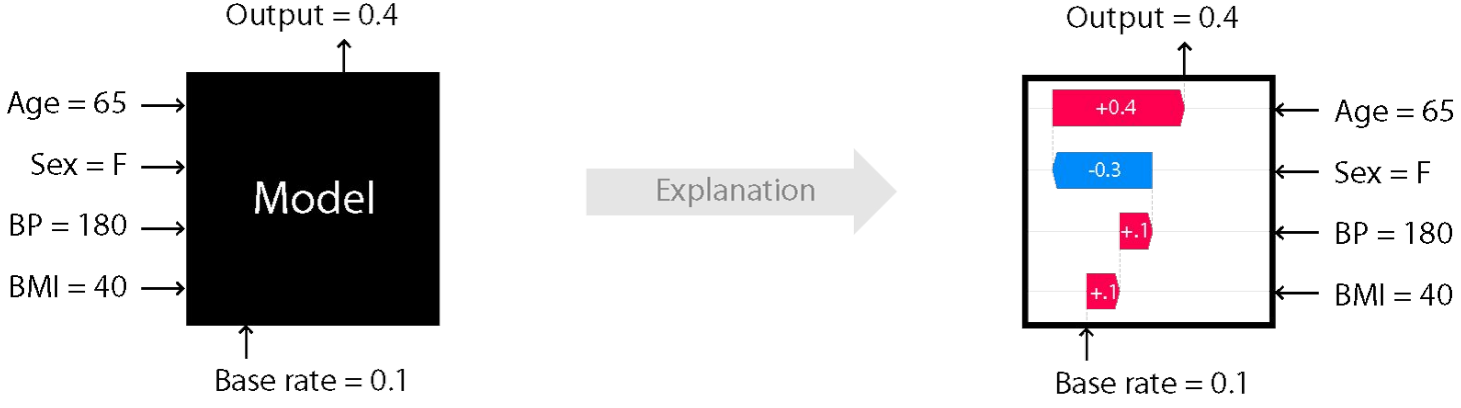


Confianza en una predicción: No podemos aceptar a ciegas la predicción de un modelo, especialmente cuando estamos decidiendo sobre la vida de otras personas..

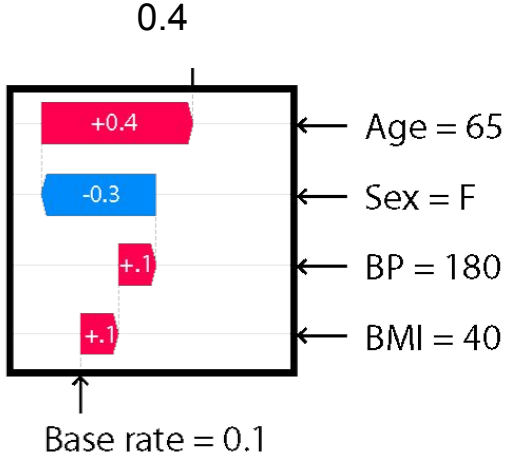
Confianza en el modelo: Más allá de las métricas, la confianza e interpretabilidad en las predicciones nos ayudan a confiar en el modelo para ponerlo en producción con datos reales.

Otro enfoque: SHAP

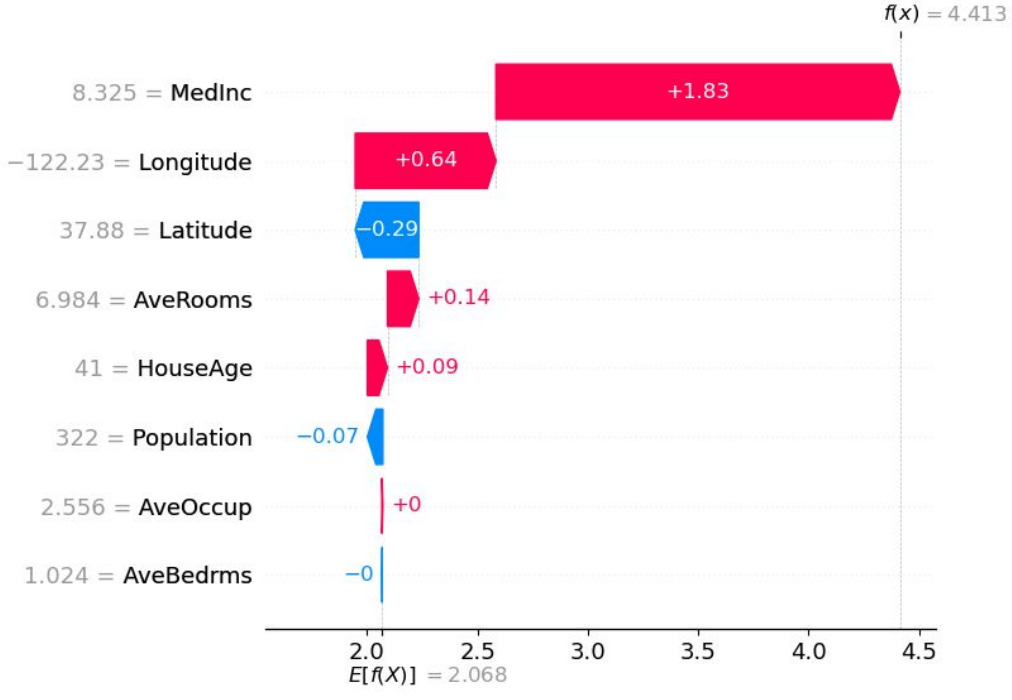
SHAP (SHapley Additive exPlanations)



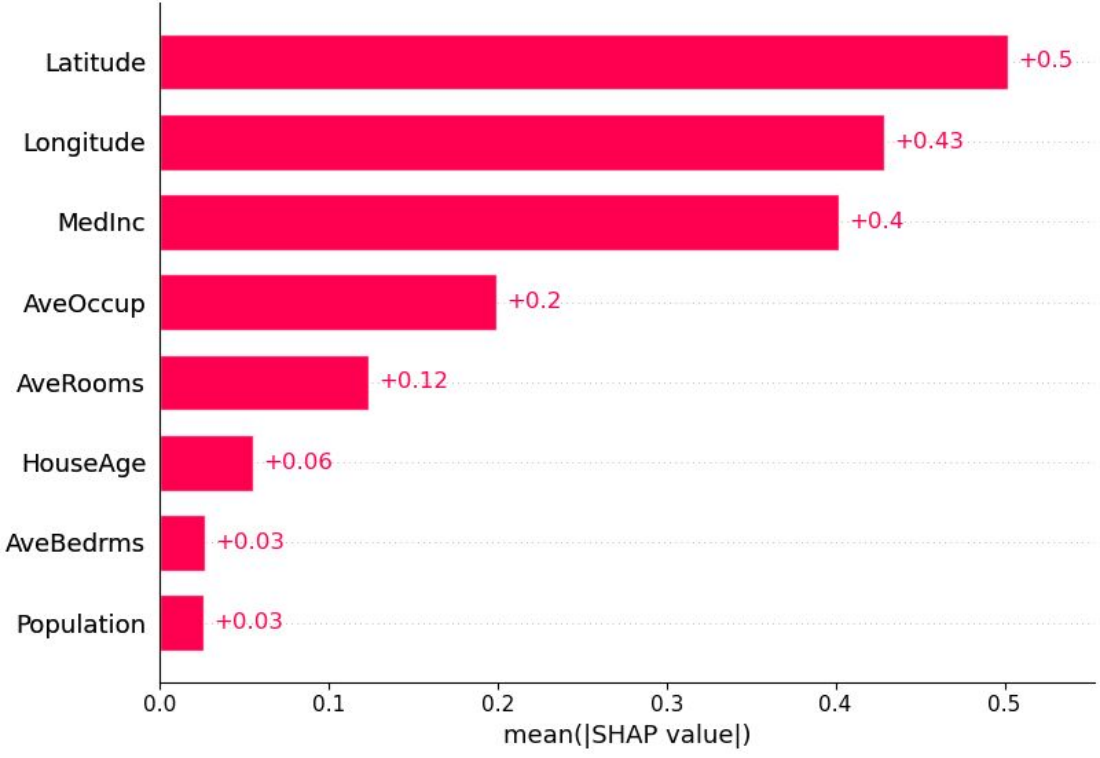
https://github.com/slundberg/shap



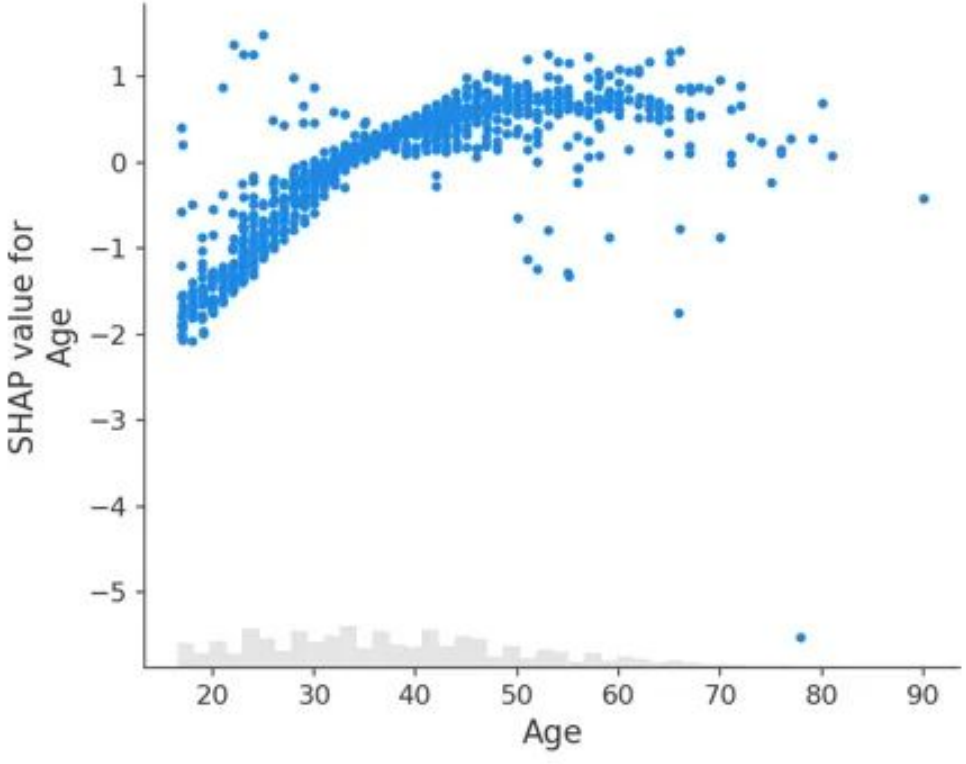
- Un método más robusto basado en el concepto de shapley value (teoría de juegos).
- Calcula la importancia para todas las permutaciones de variables y hace un promedio, esto retiene una mayor complejidad que LIME.

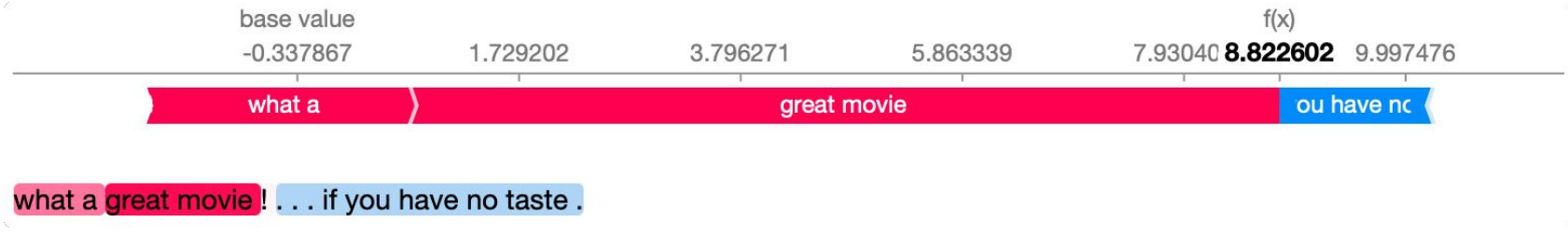

f(x) = 4.413
8.325 = MedInc
−122.23 = Longitude
37.88 = Latitude
6.984 = AveRooms
41 = HouseAge
322 = Population
2.556 = AveOccup
1.024 = AveBedrms
+1.83
+0.64
−0.29
+0.14
+0.09
−0.07
+0
−0
2.0
2.5
3.0
3.5
4.0
4.5
E[f(X)] = 2.068

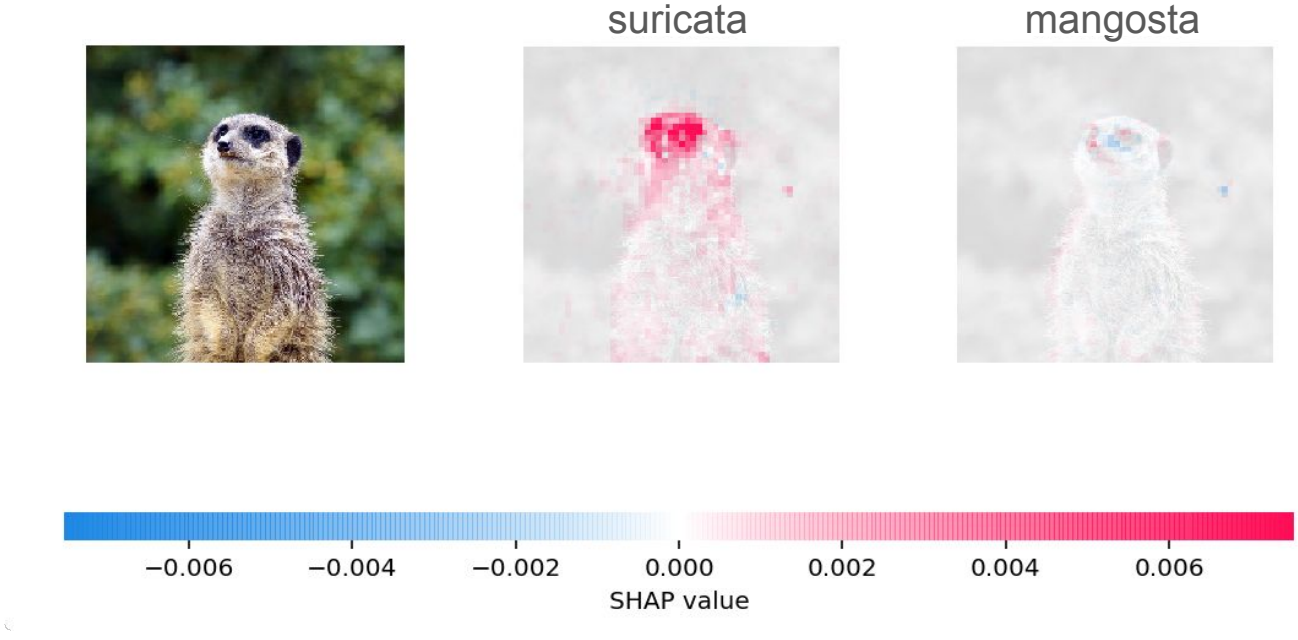
SHAP (SHapley Additive exPlanations) GLOBAL



Income vs. Age




base value
-0.337867
1.729202
3.796271
5.863339
7.93040
f(x)
8.822602
9.997476
what a
great movie
ou have nc
what a great movie ! . . . if you have no taste .


suricata
mangosta
−0.006
−0.004
−0.002
0.000
0.002
0.004
0.006
SHAP value