



Explicabilidad local, ¿cómo  
interpreto la predicción?

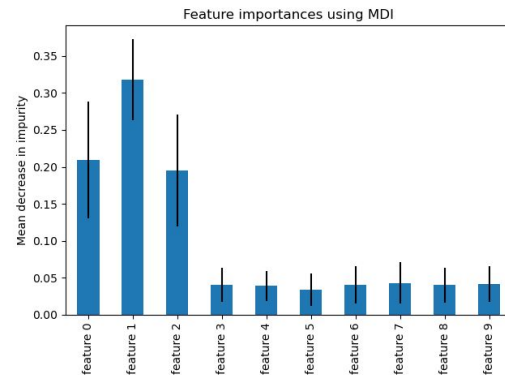
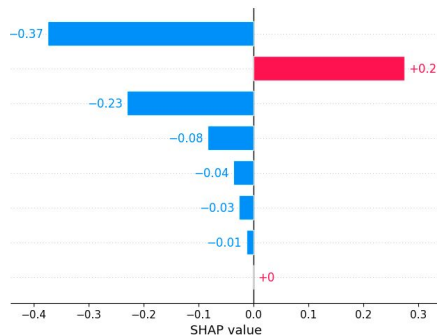
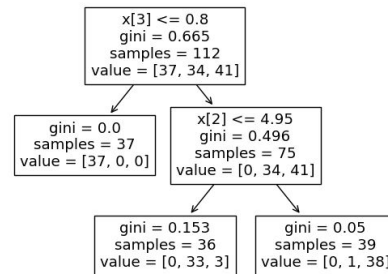


**A** ver qué es esto de la  
explicabilidad.



# Modelos explicables.

# Construidos para ser interpretados.



# Modelos explicables.

## Expli-qué.



### Model Exploration Stack

What is the model prediction for the selected instance?

$f(x)$   
AUC  
RMSE

How good is the model?

ROC curve  
LIFT, Gain charts  
Chapter 15

Which variables contribute to the selected prediction?

Break Down  
SHAP, LIME  
Chapters 6, 7, 8, 9

Which variables are important to the model?

Permutational  
Variable Importance  
Chapter 16

How does a variable affect the prediction?

Ceteris Paribus  
Chapters 10, 11

How does a variable affect the average prediction?

Partial Dependence Profile  
Accumulated Local Effects  
Chapters 17, 18

Does the model fit well around the prediction?

Chapter 12

Does the model fit well in general?

Chapter 19

PREDICTION LEVEL  
LOCAL EXPLANATIONS

MODEL LEVEL  
GLOBAL EXPLANATIONS

Explanatory Model Analysis



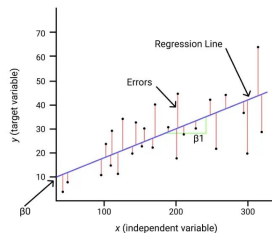
# El abanico de posibilidades.

Caja blanca.

P. ej:

Generalised Linear

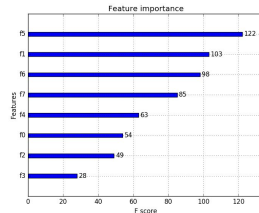
Models



Explicable globalmente.

P. ej:

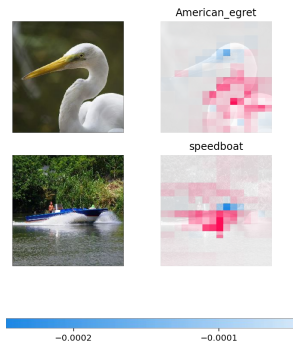
XGBoost



Explicable post-hoc.

P. ej:

Neural Networks





**B**e SHAP, my friend.



# SHapley Additive exPlanati ons.

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i,$$

where  $z' \in \{0, 1\}^M$ ,  $M$  is the number of simplified input features, and  $\phi_i \in \mathbb{R}$ .

- 01 — Basados en Teoría de juegos colaborativa.
- 02 — **Aditivos**
- 03 — Local accuracy
- 04 — Missingness
- 05 — Consistency

SHAP



# Epidemiologic Follow Up Study (NHEFS).

Personas entre 25-74 años de edad que han completado el examen médico NHANES I

	Non-Null Count	Dtype
sex_isFemale	14264	bool
age	14264	int64
physical_activity	14264	int64
...	...	...
creatinine_isUnacceptable	14264	bool
bmi	14264	float64

*dmlc*  
**XGBoost**





## El objeto principal de SHAP

# explainer.

```
explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(X)
explanations = explainer(X)
```

```
--Explainer dataframe--
.values =
array([[ 1.84628233e-01, -5.04557610e-01, -3.13231274e-02, ...,
        -1.92297876e-01, -1.64051000e-02, -5.78483492e-02],
       [ 1.66194662e-01, -1.09421897e+00,  5.98950498e-02, ...,
        -2.91730892e-02, -6.48942171e-03, -4.91099022e-02],
       [-1.49643376e-01, -1.56533003e+00,  7.20566958e-02, ...,
        -1.84312284e-01,  3.18674967e-02, -3.89468074e-02],
       ...,
       [-2.05180049e-01, -9.24388349e-01,  6.81497306e-02, ...,
        -2.10007086e-01,  2.53346898e-02,  3.22342068e-01],
       [-2.38794148e-01,  3.31081092e-01, -2.36861296e-02, ...,
        2.89917737e-01, -2.62622605e-04, -6.56308085e-02],
       [ 1.60021782e-01, -1.59773672e+00, -2.64430381e-02, ...,
        1.45278588e-01, -1.00701945e-02,  1.04265157e-02]], dtype=float32)

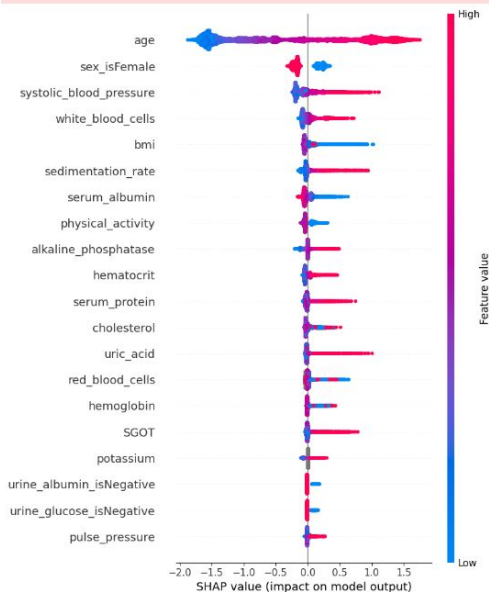
.base_values =
array([-0.6702478, -0.6702478, -0.6702478, ..., -0.6702478, -0.6702478,
       -0.6702478], dtype=float32)

.data =
array([[False, 51, 3, ..., 110.0, 40.0, 25.406802871255213],
       [False, 41, 2, ..., 136.0, 54.0, 24.58833108784943],
       [True, 31, 2, ..., 110.0, 24.0, 23.75650236105149],
       ...,
       [True, 47, 2, ..., 108.0, 28.0, 17.91519880435943],
       [True, 61, 8, ..., 180.0, 68.0, 27.33162170820975],
       [False, 29, 3, ..., 152.0, 60.0, 33.16143518222725]], dtype=object)
<class 'shap._explainer.Explainer'>
```

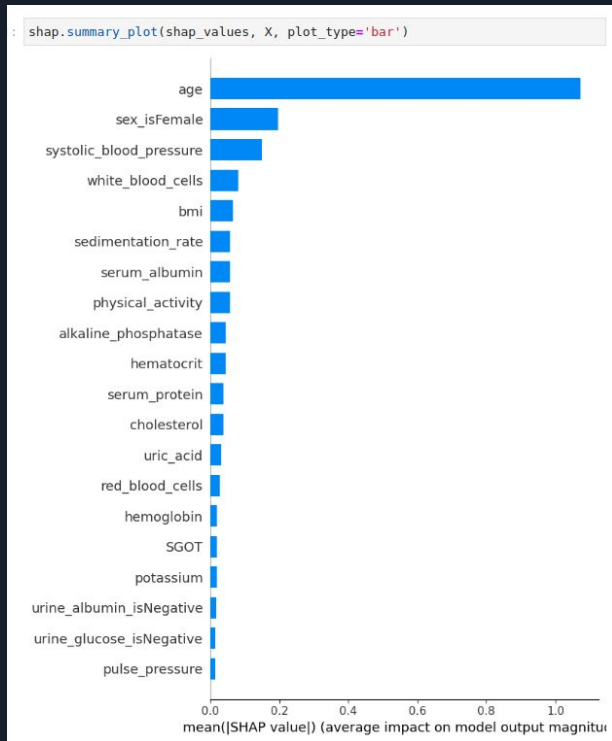
```
: shap.summary_plot(shap_values, X)

/home/bhernandez/Documents/pydata2023/venv/lib/python3.8/site-packag
es/shap/plots/_beeswarm.py:664: UserWarning:

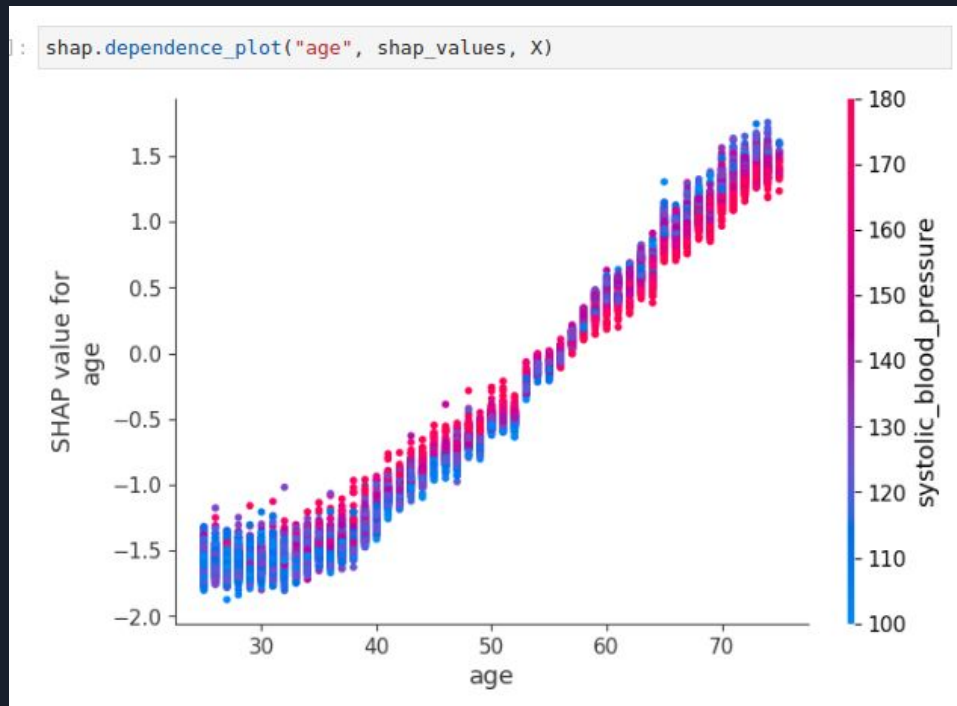
No data for colormapping provided via 'c'. Parameters 'vmin', 'vmax'
will be ignored
```



# Contribución total.



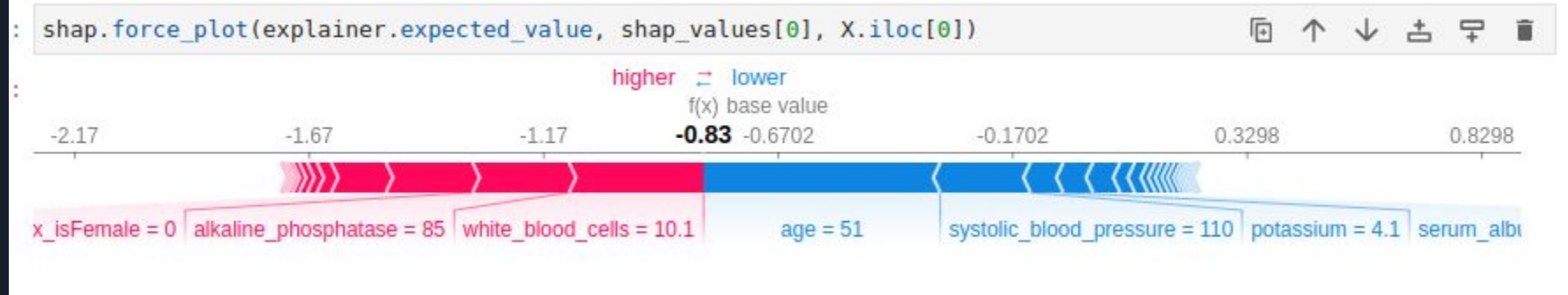
# Contribución parcial.



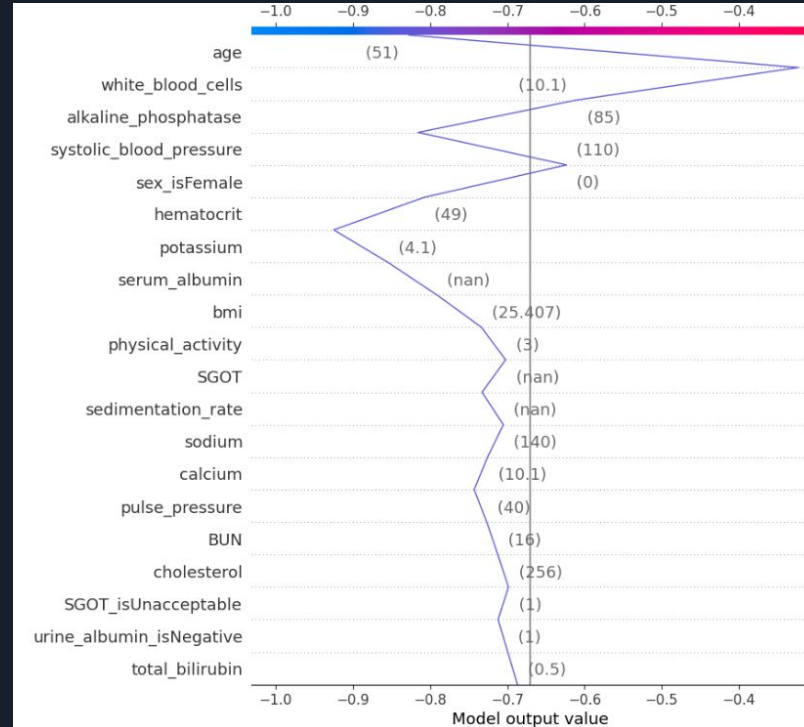


C oncretando.

# Explicabilidad local.



# Explicabilidad local.





**D** ALEX.

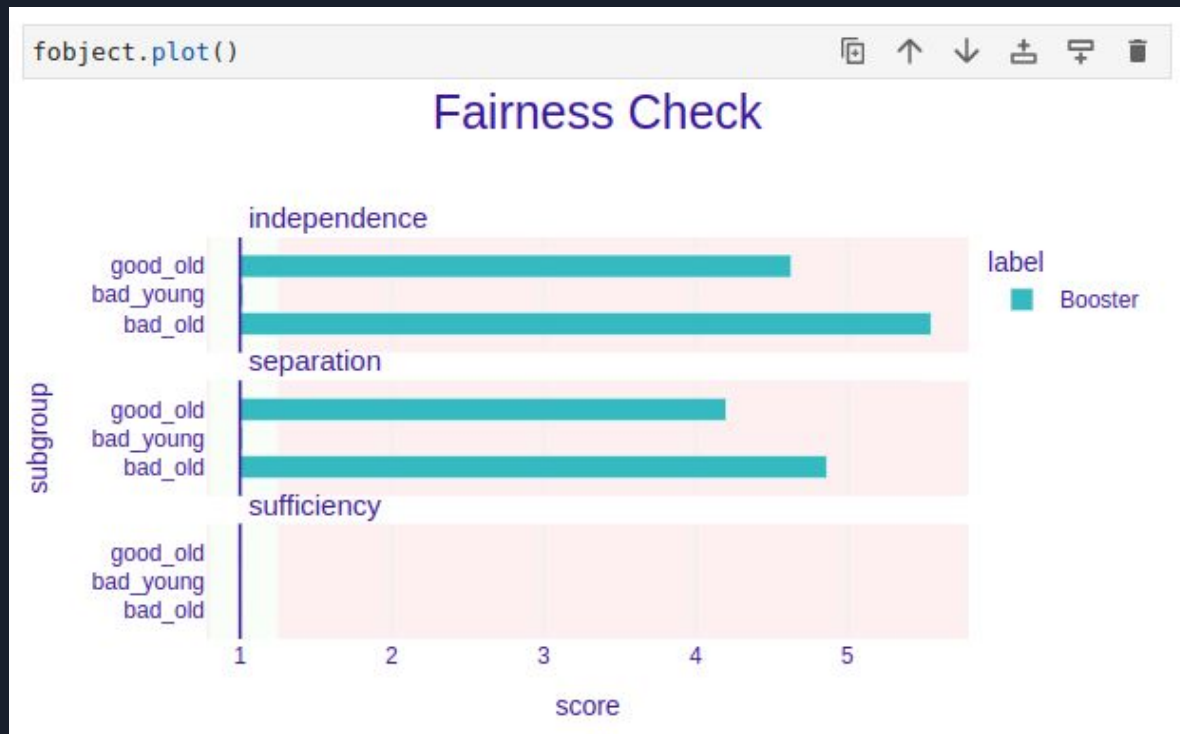


# dEscriptive mAchine Learning EXplanations.

$$\forall i \in \{a, b, \dots, z\} \epsilon < \frac{metric_i}{metric_{privileged}} < \frac{1}{\epsilon}$$



# fairness.





¡Muchas  
gracias!