XAI para la Comprensión de Sesgos:

Metodologías y herramientas

Javier Fumanal-Idocin¹
Javier Andreu-Perez¹
Raquel Fernandez-Peralta²

¹School of Computer Science and Electronic Engineering, University of Essex ²Mathematical Institute, Slovak Academy of Sciences



¿Quiénes somos?



NextGenerationEU Postdoc

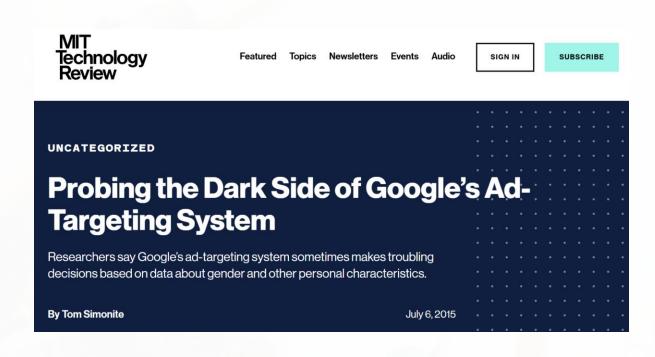


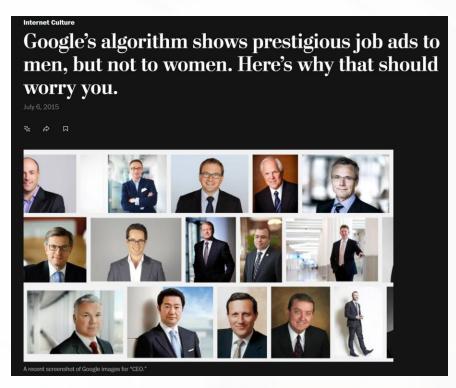


YUFE4Postdoc

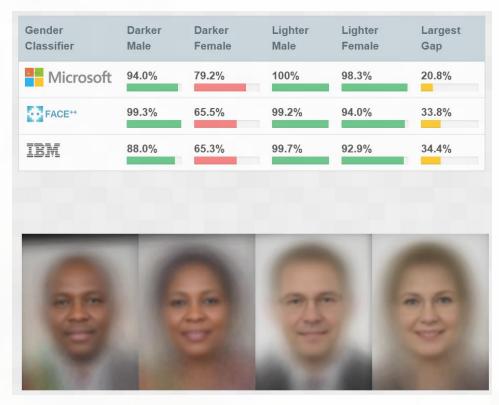


Utilizaremos el término "sesgo" para referirnos en sentido amplio al trato desigual, desfavorable e injusto de un grupo respecto de otro.





https://www.technologyreview.com/2015/07/06/110198/probing-the-dark-side-of-googles-adtargeting-system/



http://gendershades.org/overview.html

Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's.



Desiree Rios/The New York Times

Eight years after a controversy over Black people being mislabeled as gorillas by image analysis software — and despite big advances in computer vision — tech giants still fear repeating the mistake.

By Nico Grant and Kashmir Hill

May 22, 2023

https://www.nytimes.com/2023/05/22/techn ology/ai-photo-labels-google-apple.html

REVIEW ARTICLE | Originally Published 19 February 2007 |



Evidence-Based Guidelines for Cardiovascular Disease Prevention in Women: 2007 Update

Lori Mosca, MD, MPH, PhD, Chair, Carole L. Banka, PhD, Emelia J. Benjamin, MD, Kathy Berra, MSN, NP, Cheryl Bushnell, MD, Rowena J. Dolor, MD, MHS, Theodore G. Ganiats, MD, ... SHOW ALL ..., and for the Expert Panel/Writing Group AUTHOR INFO & AFFILIATIONS

Circulation • Volume 115, Number 11 • https://doi.org/10.1161/CIRCULATIONAHA.107.181546

Perspective | Open access | Published: 21 November 2024

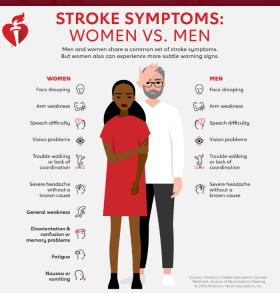
Artificial intelligence bias in the prediction and detection of cardiovascular disease

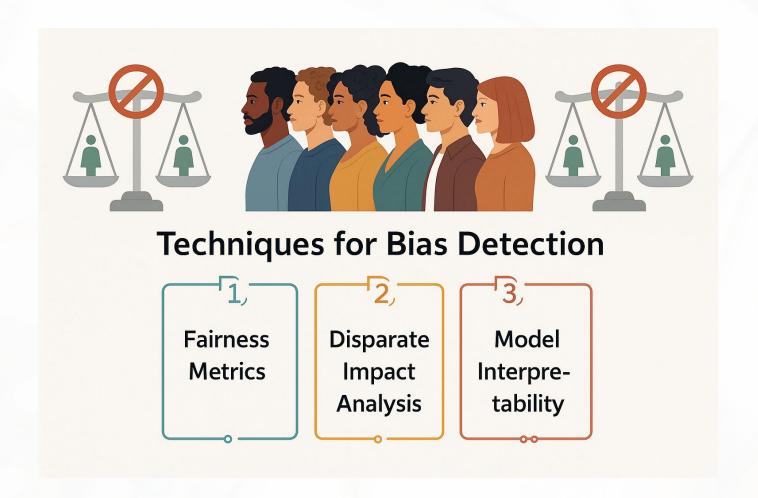
Ariana Mihan, Ambarish Pandey & Harriette G. C. Van Spall □

npj Cardiovascular Health 1, Article number: 31 (2024) | Cite this article

3494 Accesses **8** Altmetric Metrics







Toolkits mitigación y detección de sesgos

Fairlearn

= Fairlearn

https://github.com/fairlearn/fairlearn

AIF360



https://github.com/Trusted-AI/AIF360

https://github.com/datamllab/awesome-fairness-in-ai

El derecho a una explicación

European Union regulations on algorithmic decision-making and a "right to explanation"

Bryce Goodman,^{1*} Seth Flaxman,²

¹Oxford Internet Institute, Oxford
 1 St Giles', Oxford OX1 3LB, United Kingdom
 ²Department of Statistics, University of Oxford,
 24-29 St Giles', Oxford OX1 3LB, United Kingdom

*To whom correspondence should be addressed; E-mail: flaxman@stats.ox.ac.uk.

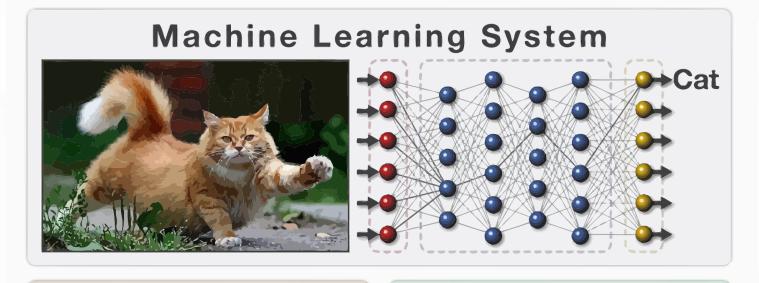
Abstract

We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine learning algorithms. Slated to take effect as law across the EU in 2018, it will restrict automated individual decision-making (that is, algorithms that make decisions based on user-level predictors) which "significantly affect" users. The law will also effectively create a "right to explanation," whereby a user can ask for an explanation of an algorithmic decision that was made about them. We argue that while this law will pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks which avoid discrimination and enable explanation.



https://arxiv.org/abs/1606.08813

Inteligencia Artificial Explicativa (XAI)



This is a cat.

Current Explanation

This is a cat:

- It has fur, whiskers, and claws.
- It has this feature:

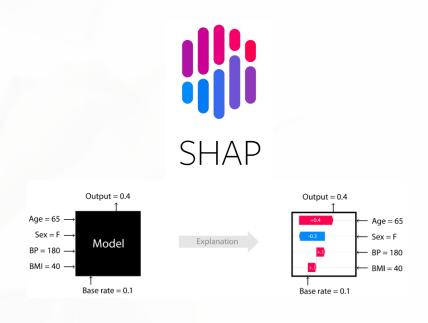




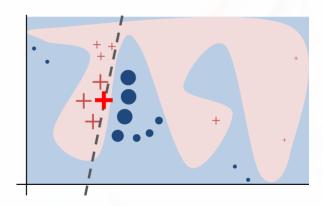
XAI Explanation

© DARPA

Construyendo explicaciones



LIME

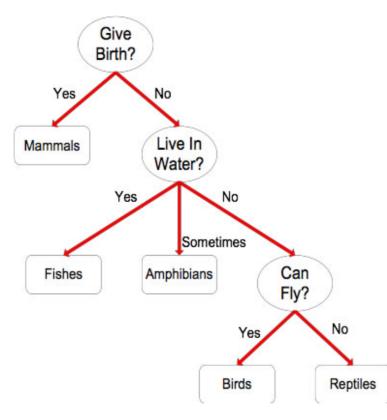


- SHAP: https://github.com/shap/shap
- LIME: https://github.com/marcotcr/lime
- OmniXAI: https://github.com/salesforce/omnixai

Clasificación con reglas

Algunos clasificadores se consideran interpretables, como aquellos que clasifican utilizando un conjunto de reglas 'si...entonces'.

C4.5 Decision Tree versus C4.5 rules versus RIPPER



C4.5rules

(Give Birth=No, Can Fly=Yes) → Birds (Give Birth=No, Live in Water=Yes) → Fishes (Give Birth=Yes) → Mammals (Give Birth=No, Can Fly=No, Live in Water=No) → Reptiles () → Amphibians

RIPPER:

(Live in Water=Yes) → Fishes (Have Legs=No) → Reptiles (Give Birth=No, Can Fly=No, Live In Water=No) → Reptiles (Can Fly=Yes, Give Birth=No) → Birds () → Mammals

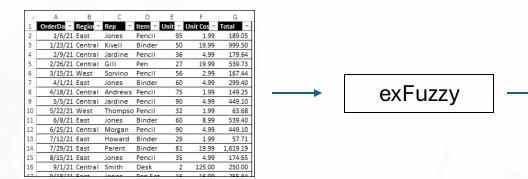
Confusion Matrix for C4.5 versus RIPPER



exFuzzy

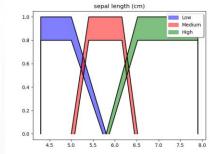
¿Qué es exFuzzy?

- Una biblioteca python para realizar XAI utilizando reglas difusas
 - Con especial atención a la explicabilidad y la visualización



Rules for consequent: setosa
IF petal width (cm) IS Low WITH DS 0.7359951780154956, ACC 0.9444444444444444444444444444444444444
Rules for consequent: versicolor
IF petal width (cm) IS Medium WITH DS 0.7630571825431511, ACC 0.90625
Rules for consequent: virginica
IF sepal length (cm) IS High AND petal width (cm) IS High WITH DS 0.5742697564048905, ACC 0.9354838709677419 IF sepal length (cm) IS Low AND petal length (cm) IS High WITH DS 0.005588278153979678, ACC 1.0

Cluster	Positivity	Negativity	Subjectivity	Polarity	DS	Acc
1	•	•	•	•	0.07	0.40
	•	•	•	•	0.12	0.6
2	•	•	•	•	0.54	0.8
	•	•	•	•	0.11	0.7
3	•	•	•	•	0.04	0.1
4	-		-	-	-	-



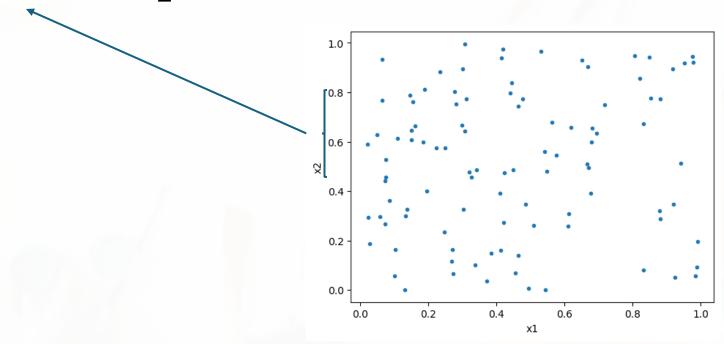
- Los métodos de clasificación estándar cuentan con numerosos recursos:
 - NN (MLP y DL)
 - SVM
 - Random Forest
- Scikit Learn sigue siendo la biblioteca de ciencia de datos más descargada!

• Los métodos de reglas son técnicas XAI basados en condicionales lógicas:

IF x_1 is A AND x_2 is B THEN do Z/class is Z

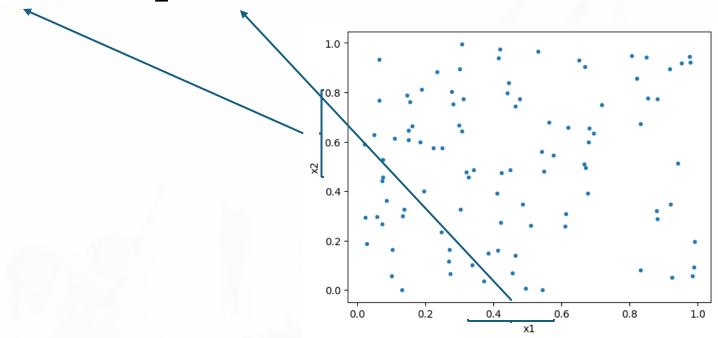
 Los métodos de reglas son técnicas XAI basados encondicionales lógicas:

IF x_1 is A AND x_2 is B THEN do Z/ class is Z



Rule-based/logic constrain is another XAI method*:

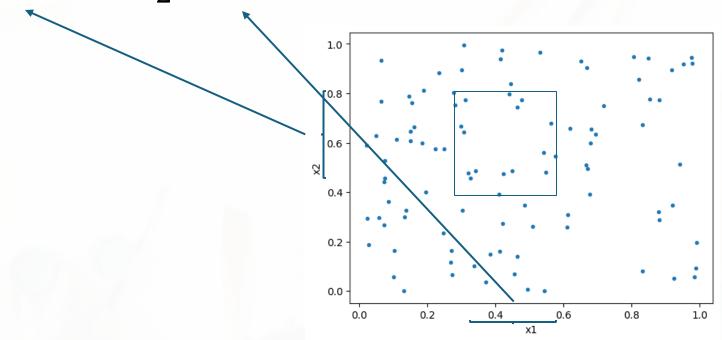
IF x_1 is A AND x_2 is B THEN do Z/ class is Z



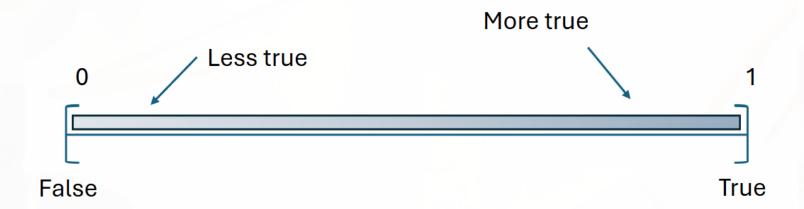
Why exFuzzy?

 Los métodos de reglas son técnicas XAI basados encondicionales lógicas:

IF x_1 is A AND x_2 is B THEN do Z/ class is Z



- Nos salimos de la lógica booleana 0/1 blanco/negro.
- Los valores de la lógica difusa son escalas de grises en el intervalo [0,1].





- Temperatura/sensación térmica: caliente/frío



- Proximidad: cerca/lejos



- Precios: barato/caro



- Arte: impresionismo/cubismo

• Las reglas difusas son un tipo particular de reglas muy útiles debido a cómo dividen el estado de entrada.



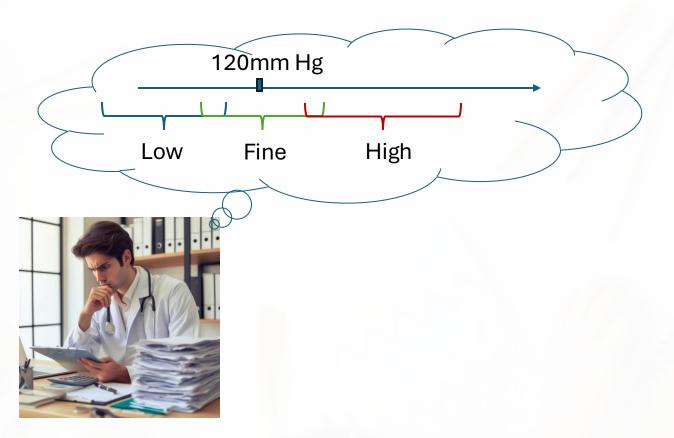
Regla clásica

If blood pressure < 100 mg

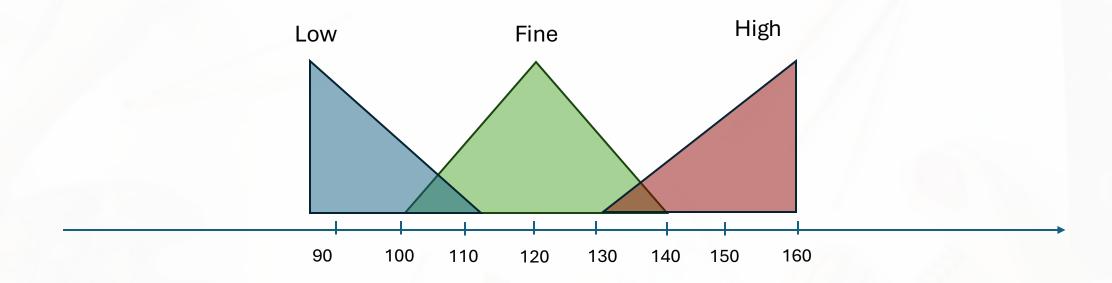
Regla Difusa

If blood pressure Low

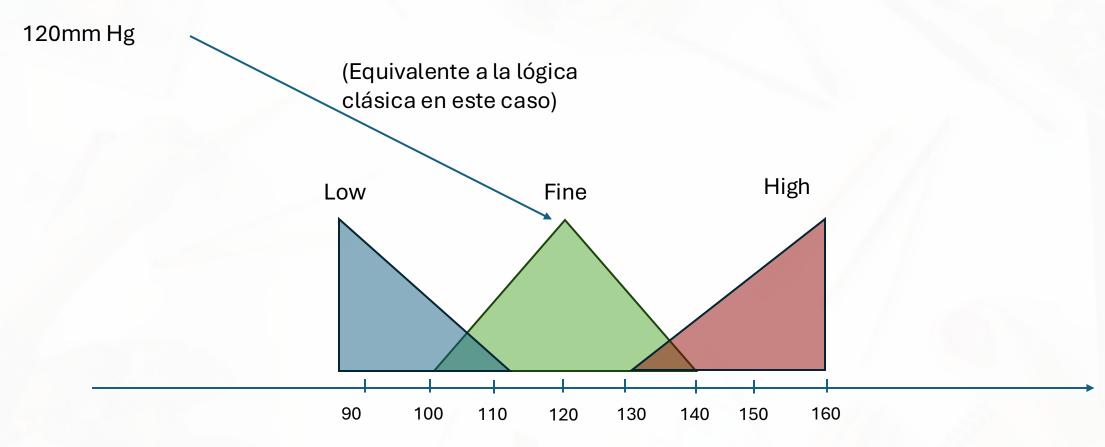
• Las reglas difusas son un tipo particular de reglas muy útiles debido a cómo dividen el estado de entrada.



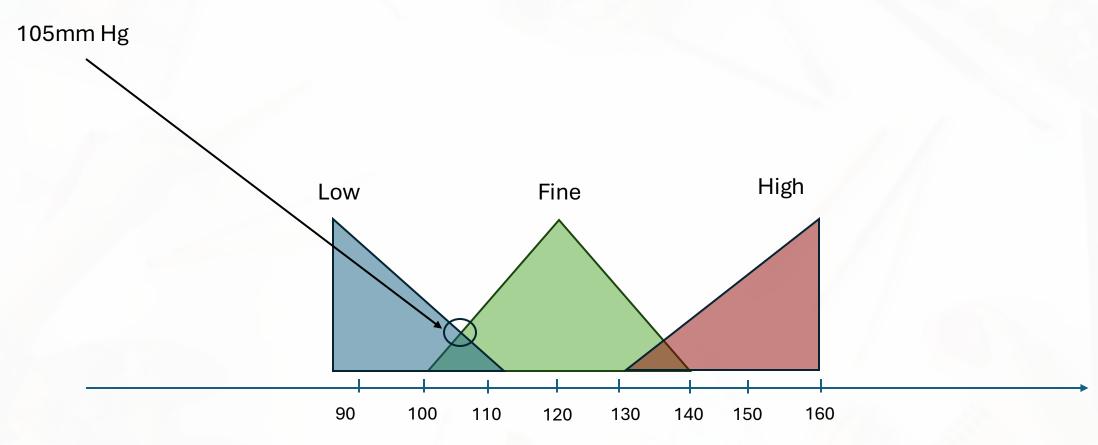
Particiones Difusas



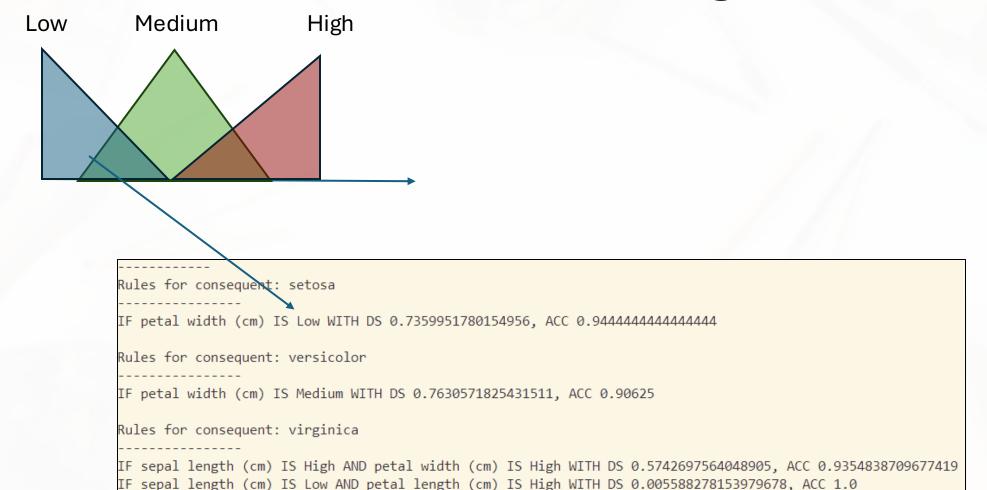
Particiones Difusas



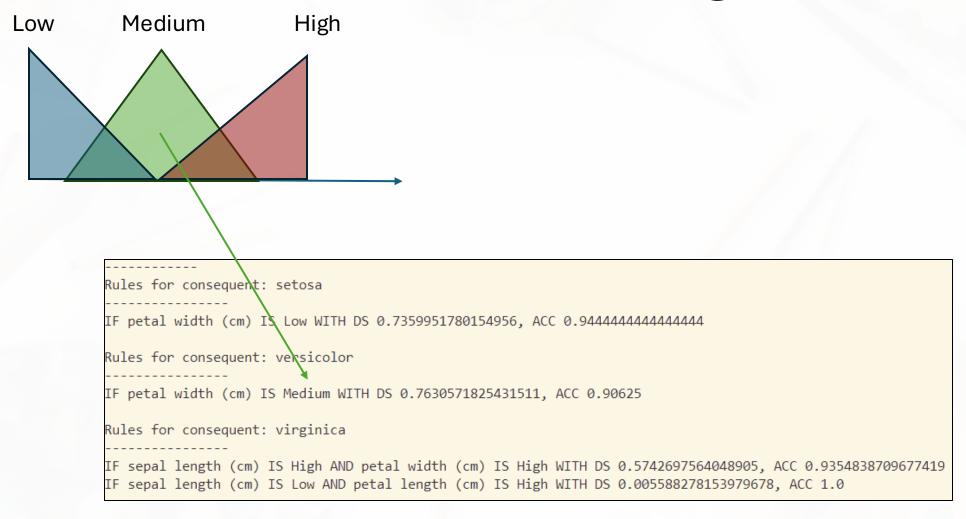
Particiones Difusas



Razonamiento con reglas



Razonamiento con Reglas



exFuzzy

- Python
- Simplicidad para los pipelines Scikit-Learn.
- Facilidad de visualización y reporte de resultados.
- FOCO en la EXPLICABILIDAD (especialmente en la forma de aprender las reglas)
- Para los entusiastas: soporte para diferentes tipos de conjuntos difusos, conjuntos difusos automatizados o precalculados, coberturas lingüísticas, etc.

Inferencia difusa clásica

- Calcular el nivel de verdad de cada regla
 - Los operadores lógicos clásicos son sustituidos por funciones. Por ejemplo, la conjunción es el producto.
- Se asigna un peso a cada regla.
- A cada ejemplo se le asigna la clase de la regla con un mayor nivel de verdad y peso.
 - Hay otras alternativas que utilizan diversas reglas para clasificar, pero se ha elegido esta opción para que la interpretación sea más sencilla.

- Seleccionamos los parámetros:
 - Máximo número de reglas
 - Máximo número de variables en el antecedente
 - Número de etiquetas en la partición difusa
 - Tipo de conjunto difuso (t1, t2)

- Optimización genética:
 - Establecemos el número de generaciones y el número de individuos.
 - Ajuste mediante el coeficiente de correlación de Matthew.
 - Aunque hay posibilidad de cambiar esta función

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$



• Código (No hay variables lingüísticas precalculadas, se optimizan con el resto de las reglas).

• Código:

```
Número máximo de
import ex_fuzzy.fuzzy_sets as fs
                                                 reglas
import ex_fuzzy.evolutionary_fit as GA
import ex_fuzzy.utils as utils
import ex fuzzy.eval tools as eval tools
iris = datasets.load iris()
X = pd.DataFrame(iris.data, columns=iris.feature names)
v = iris.target
X train, X test, y train, y test = train test spl/it(X, y, test size=0.33, random state=0)
fl classifier = GA.BaseFuzzyRulesClassifier(hRules=10, nAnts=4, n_linguist_variables=3,
                                            fuzzy type=fs.FUZZY SETS.t2, tolerance=0.001)
fl classifier.fit(X train, y train, n gen=50, pop size=30)
eval tools.eval fuzzy model(fl classifier, X train, y train, X test, y test,
                       plot rules=True, print rules=True, plot partitions=True)
```

• Código:

Número máximo de condiciones en

• Código:

```
Número de particiones en variables numéricas

import ex_fuzzy.evolutionary_fit as GA
import ex_fuzzy.utils as utils
import ex_fuzzy.eval_tools as eval_tools

iris = datasets.load_iris()
X = pd.DataFrame(iris.data, columns=iris.feature_names)
y = iris.target

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=0)
fl_classifier = GA.BaseFuzzyRulesClassifier(nRules=10, nAnts=4, n_linguist_variables=3, fuzzy_type=fs.Fuzzy_sets.tz, tolerance=0.001)
fl_classifier.fit(X_train, y_train, n_gen=50, pop_size=30)

eval_tools.eval_fuzzy_model(fl_classifier, X_train, y_train, X_test, y_test, plot_rules=True, print_rules=True, plot_partitions=True)
```

• Código:

```
import ex_fuzzy.fuzzy_sets as fs
import ex_fuzzy.evolutionary_fit as GA
import ex_fuzzy.utils as utils
import ex_fuzzy.eval_tools as eval_tools

iris = datasets.load_iris()
X = pd.DataFrame(iris.data, columns=iris.feature_names)
y = iris.target

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=0)
fl_classifier = GA.BaseFuzzyRulesClassifier(nRules=10, nAnts=4, n_linguist_variables=3, fuzzy_type=fs.FUZZY_SETS.t2, tolerance=0.001)
fl_classifier.fit(X_train, y_train, n_gen=50, pop_size=30)
eval_tools.eval_fuzzy_model(fl_classifier, X_train, y_train, X_test, y_test, plot_rules=True, print_rules=True, plot_partitions=True)
```

Entrenando un clasificador con exFuzzy

• Código:

```
import ex_fuzzy.fuzzy_sets as fs
import ex_fuzzy.evolutionary_fit as GA
import ex_fuzzy.utils as utils
import ex_fuzzy.eval_tools as eval_tools

iris = datasets.load_iris()
X = pd.DataFrame(iris.data, columns=iris.feature_names)
y = iris.target

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=0)
fl_classifier = GA.BaseFuzzyRulesClassifier(nRules=10, nAnts=4, n_linguist_variables=3, fuzzy_type=fs.FUZZY_SETS.t2, tolerance=0.001)
fl_classifier.fit(X_train, y_train, n_gen=50, pop_size=30)

eval_tools.eval_fuzzy_model(fl_classifier, X_train, y_train, X_test, y_test, plot_rules=True, print_rules=True, plot_partitions=True)
```

Visualizando las reglas resultantes

- El algoritmo te devuelve la base de reglas obtenida:
 - También se obtiene la dominancia/precisión de cada regla

Reglas fáciles de entender y guardar

```
Rules for consequent: setosa
                                                                                                                                                     File
IF petal width (cm) IS Low WITH DS 0.7359951780154956, ACC 0.944444444444444444
Rules for consequent: versicolor
IF petal width (cm) IS Medium WITH DS 0.7630571825431511, ACC 0.90625
Rules for consequent: virginica
IF sepal length (cm) IS High AND petal width (cm) IS High WITH DS 0.5742697564048905, ACC 0.9354838709677419
                                                                                                                                              Rules in text
IF sepal length (cm) IS Low AND petal length (cm) IS High WITH DS 0.005588278153979678, ACC 1.0
                                                                                                                                                   format
         * Load rules from a plain text file
        with open('iris_rules.txt', 'r') as f:
            str rules = f.read()
         Persistence of the rules example
        mrule_base = <mark>persistence.l</mark>oad_fuzzy_rules(str_rules, precomputed_partitions)
        fl classifier2 = GA.BaseFuzzyRulesClassifier(precomputed rules=mrule base)
         # fl classifier2.load master rule base(mrule base) # (Another possibility)
```

Exportar resultados para artículos

• Soporta tablas de latex

Author		Antecedents	DS	Train Acc	Test Acc
	1	IF New Realism IS Low AND Post Impressionism IS Medium	0.0076	0.5000	0.0000
	2	IF Early Renaissance IS Medium AND New Realism IS Medium AND Synthetic Cubism IS Medium	0.0740	0.7777	1.0000
	3	IF Early Renaissance IS Low AND Synthetic Cubism IS High	0.2517	0.9390	0.8888
	4	IF Synthetic Cubism IS Low	0.4624	0.9097	0.9250
	5	IF Contemporary Realism IS Medium AND Synthetic Cubism IS Low AND Relevant area IS Low	0.0092	0.0000	0.0000
- Park Inc. Mar.					
E	6	IF Contemporary Realism IS Medium AND Minimalism IS Low	0.3389	0.7586	0.7692
	7	IF Early Renaissance IS Medium AND Minimalism IS Medium AND Synthetic Cubism IS Medium	0.0124	0.0000	0.0000

Fumanal-Idocin, J., Andreu-Perez, J., Cord, O., Hagras, H., & Bustince, H. (2023). Artxai: Explainable artificial intelligence curates deep representation learning for artistic images using fuzzy techniques. *IEEE Transactions on Fuzzy Systems*.

¿Porqué exFuzzy?

Librerías "difusas":



(Python)

• Scikit-fuzzy, Simpful



(Java)

(Java)
• Juzzy (many more)



(MATLAB)

Fuzzy logic toolbox





• FuzzyLogic.jl, fuzzy.jl

¿Porqué exFuzzy?

• Librerías "difusas":



(Python)

• Scikit-fuzzy, Simpful



(Java)

Juzzy (many more)



(MATLAB)

• Fuzzy logic toolbox



(R)

Lfl



• FuzzyLogic.jl, fuzzy.jl



¿Porqué exFuzzy?

• Librerías "difusas":



(Python)

• Scikit-fuzzy, Simpful



(Java)

• Juzzy (many more)



(MATLAB)

Fuzzy logic toolbox



(R)

Lfl



• FuzzyLogic.jl, fuzzy.jl

Sin módulos de aprendizaje

Solo ANFIS

Requiere familiaridad con la teoría de la lógica difusa



https://github.com/Fuminides/ex-fuzzy

Instalar exFuzzy

- Desde pypi: pip install ex-fuzzy
- Desde github:
 - > git clone https://github.com/Fuminides/ex-fuzzy.git
 - > cd ex-fuzzy
 - > pip install.



Tutorial: Detección y mitigación de sesgos con exFuzzy

https://github.com/Fuminides/WorkshopIgual dad2025

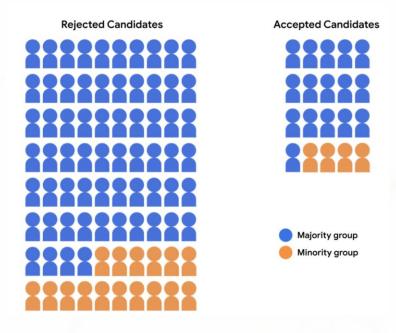
Ejemplo 1 - Titanic



Ejemplo 2 – Supervivencia a fallo cardiaco



Ejemplo 3 – Concesión de préstamos





Resumen Final

En este workshop hemos visto...

- Como los algoritmos de XAI nos pueden ayudar a identificar posibles sesgos en nuestros resultados.
- Introducción a la librería exFuzzy como un recurso más para entrenar clasificadores basados en reglas, los cuales son interpretables.
- Algunas técnicas sencillas para mitigación de sesgos, que pueden utilizar conjuntamente con la librería.

Algunos proyectos futuros

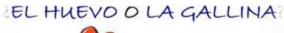
• Cuantificación de incertidumbre.

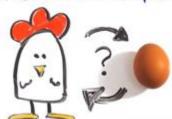




• Descubrimiento de subgrupos

Direccionalidad





¡Gracias!

Universidad Carlos III, Madrid, 2025



Lee la documentación!

https://fuminides.github.io/ex-fuzzy/



Usa el código!

https://github.com/Fuminides/ex-fuzzy



Cita el artículo!

https://www.sciencedirect.com/science/article/pii/S0925231224008191









