CALIME: Causality-Aware Local Interpretable Model-Agnostic Explanations

Martina Cinquini, Riccardo Guidotti

Computer Science Department, University of Pisa, Italy

SCAN ME

OVERVIEW

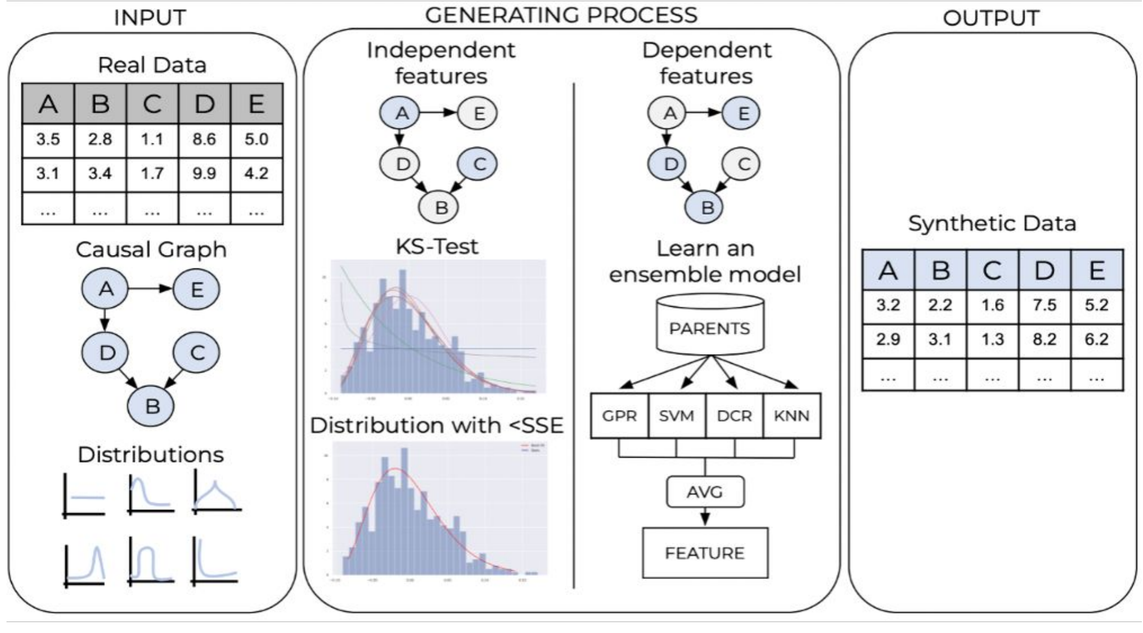
MOTIVATION

XAI approaches **do not take into account** causal relations among input features

OUR CONTRIBUTION

LIME variant **incorporates causal links in the explanation extraction process**

GENCDA

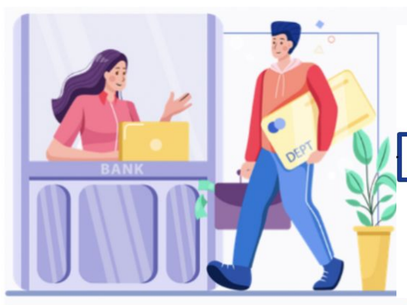


WHY DO WE NEED CAUSALITY?

Goal: Can the customer get the loan?

Black Box Prediction: No, the loan is denied

LIME Explanation: Low education level is mainly responsible

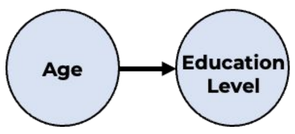


Age	Income	Education Level	Weekly working hours
24	800	High School	20
28	1300	Bachelor Degree	35
...	...	...	...

Generated Neighborhood

24	800	PHD	20

The generated instance is not plausible

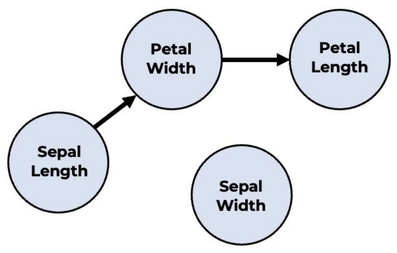


NEIGHBORHOOD GENERATION

Instance to explain

Petal Length	Sepal Length	Petal Width	Sepal Width
4.9	3.0	1.4	0.2

Causal Graph



Permutations based on the Causal Graph

Change feature values of source nodes and isolated nodes

Petal Length	Sepal Length	Petal Width	Sepal Width
4.9	**3.4**	1.4	**0.5**

Change feature values of parent node and, subsequently, their child(s)

Petal Length	Sepal Length	Petal Width	Sepal Width
5.3	**3.4**	**1.2**	0.2

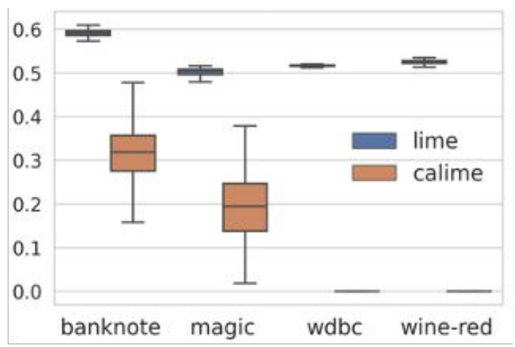
EXPERIMENTS

Datasets

banknote, wdbc, magic, wine-red

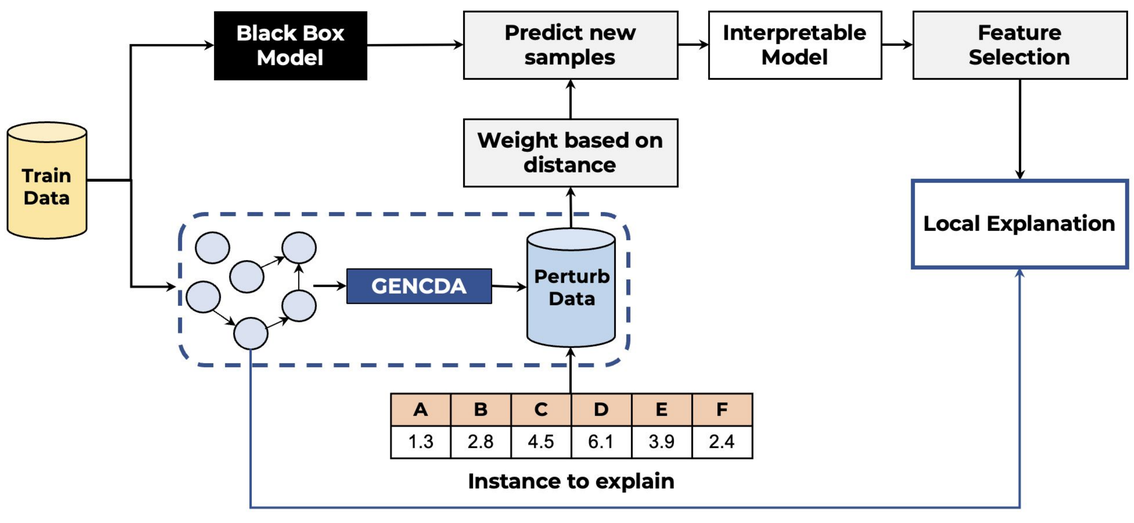
Evaluation metrics

Fidelity, Plausibility, Stability



CALIME outperforms LIME in both **black-box fidelity** and **explanations plausibility**

OUR FRAMEWORK



KEY TAKEAWAY

CALIME is the **first approach** able to **infer** and **integrate causal relations** to **promote interpretability** of Machine Learning models