Local model-agnostic explanations

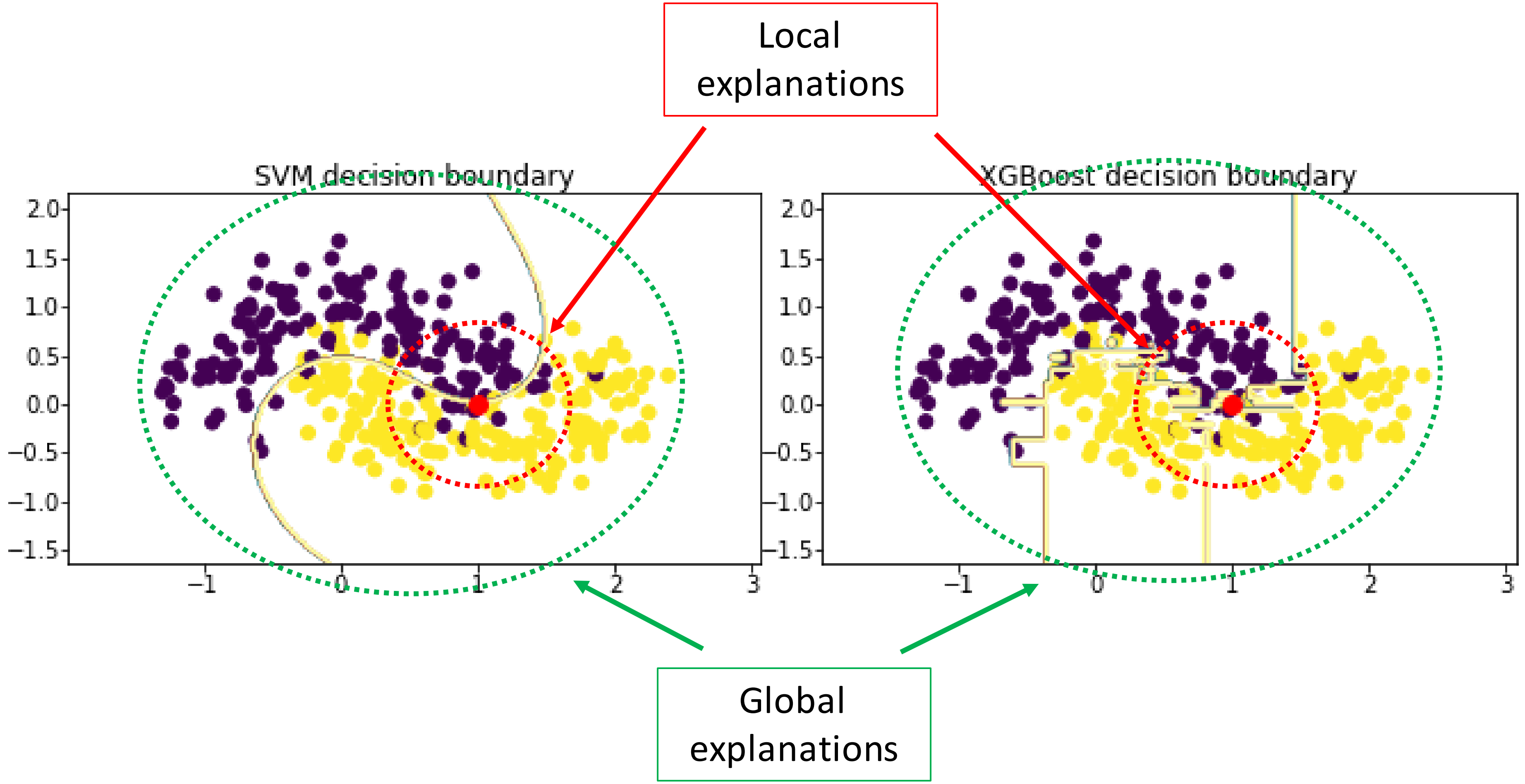
Szymon Bobek

Jagiellonian University
2023

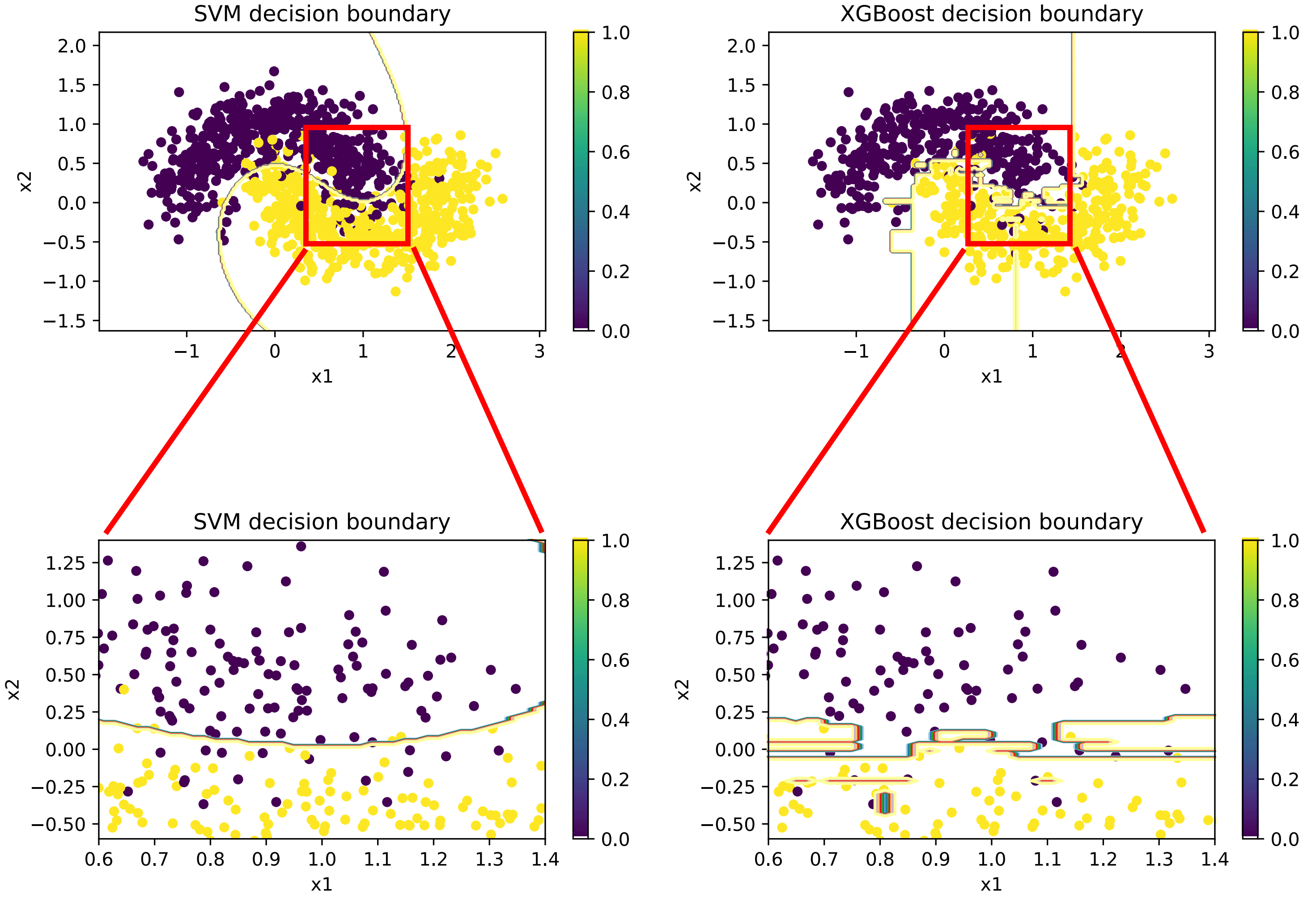
JAGIELLONIAN UNIVERSITY
IN KRAKÓW

https://geist.re

Local vs Global explanations



Locally, the decision boundary is simpler



- In this approach we focus on explining an instance
- "Zooming in" we can fit inherently interpretable model that will approximate the decision of the blackbox one
- The assumption is not always valid. There are models which has complex decision boundary even locally
- Term "Locally" is vague. The locallity is subjective
- When zooming in, we are limiting the number of samples that can be used for training
- What in case of instances that are far from the distribution?

Local Model-Agnostic Surrogate Model

Locality-aware loss function

Balc-box model

Local surrogate model

$$\phi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Interpretable model

Size of the neiborhood (locality)

Complexity of interpretable model

Why should I trust you?

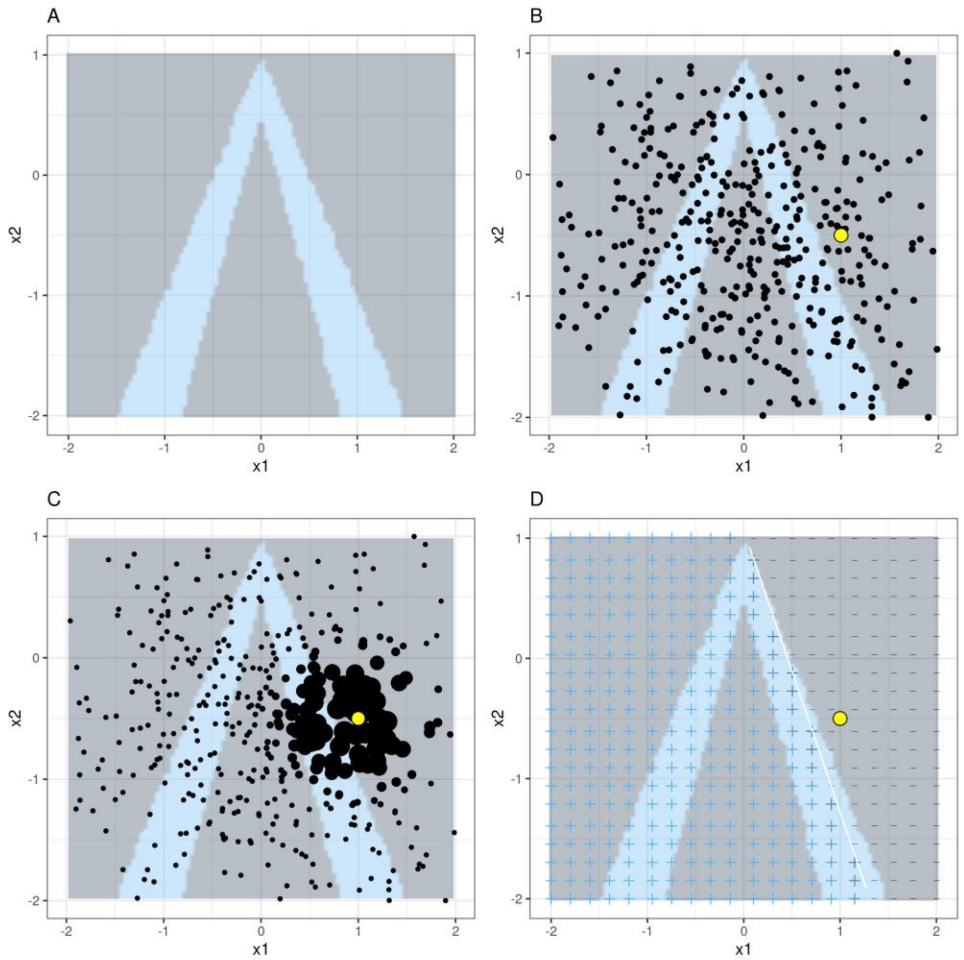
MSE for regression

Balc-box model

$$\phi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Linear regression
Lasso

Exponential
smoothing Kernel



Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

Anchors: High-Precision Model-Agnostic Explanations

Anchors

Precission and coverage

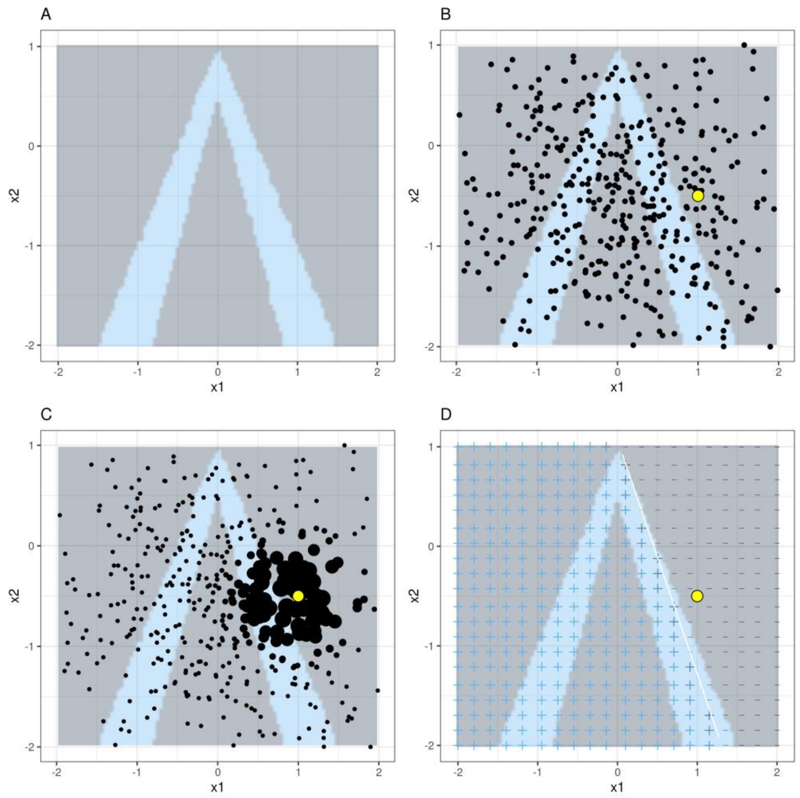
Balc-box model

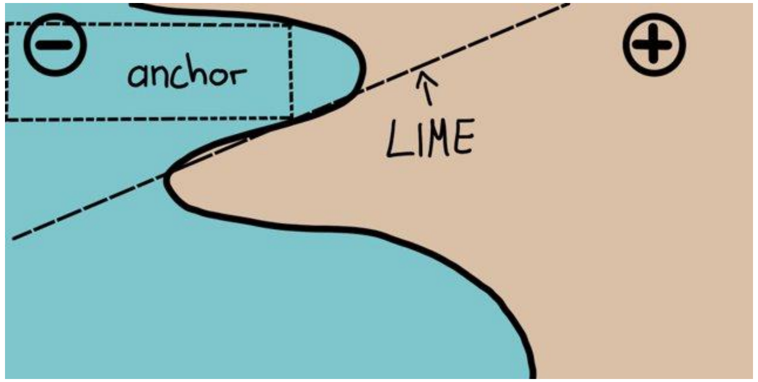
Complexity limited by the length of the rule

$$\phi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Scoped rules

Neighborhood: defined by the desired precission and anchor A

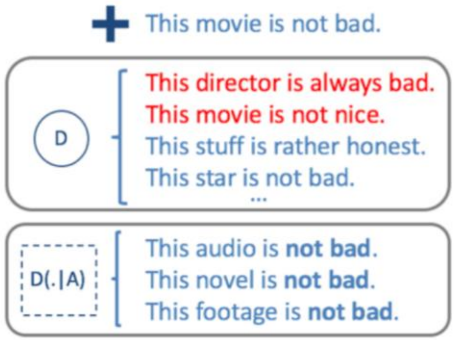


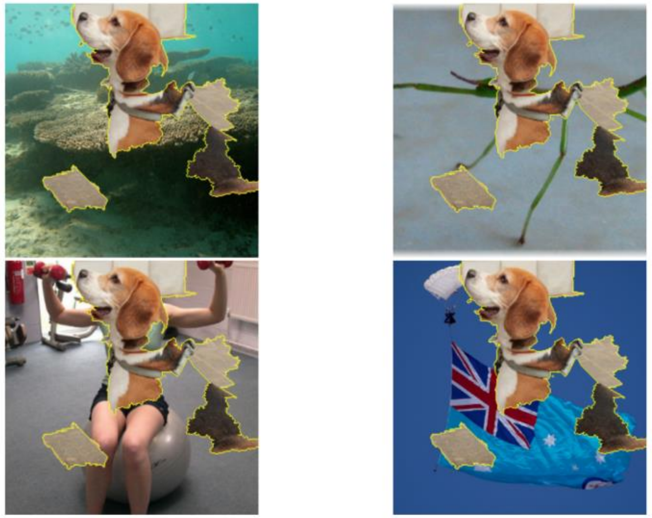


Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.11491

Perturbation

A is an anchor rule. Basically a conjunction of conditions. All possible combinations of values/operators are generated for each candidate





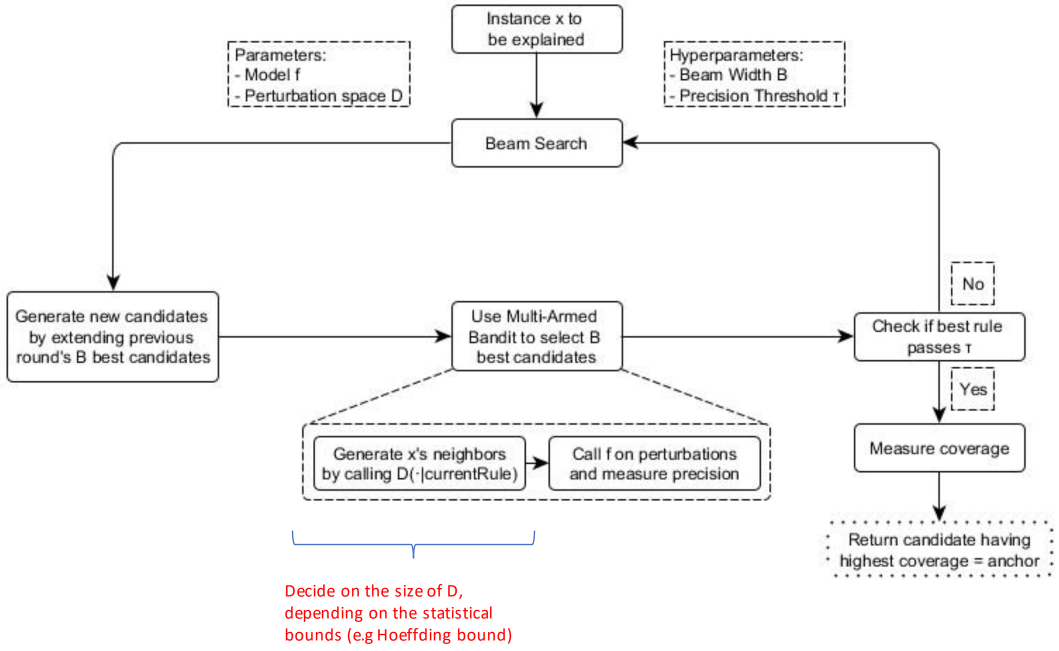
$$\mathbb{E}_{\mathcal{D}_x(z|A)}[1_{\hat{f}(x)=\hat{f}(z)}] \geq \tau, A(x) = 1$$

$D_x(z|A)$ is a perturbation datset, where the rule/amchor *A* holds

Threshold precision that is desired to be achieved by the anchor

- Anchor is also based on the (similar) perturbation idea as LIME.
- For tabular data, the features of *A* are fixed, and the rest of the row is sampled as a whole.
- For text, we reaplace other tokens with similar tokens from embedding space
- For image, we replace missing superpixels not by graying them out, but by putting some random image in a background

Anchors creation




```
BBox prediction: >50K

Anchor: Education = Bachelors AND Relationship = Husband AND
Occupation = Sales
Precision: 0.97 Coverage: 0.02
```

- Select instance to be explained and geenrate candidate anchors over perturbed dataset with Beam Search (to evaluate smarter than one +1 candidate at a time)
- Evaluate candidates with MAB approach to reduce number of calls to the model (each A is an arm)
- Extend the candidates by additional predicate
- If precision passes the threshold, return anchor

Pros and cons

- Advantages
  - Produces rules with desired precission and coverage
  - Works for all types of data modality
  - Model-agnostic
  - Non-linear as opposed to LIME

- Disadvantages
  - Highly parametrized due to MBA and Beam Search algorithms used as its core components
  - Does not produce counterfactuals
  - The rules can be very specific (long and unintuitive)
  - It is based on data generation, therefore, may create anchor that is bounded over regions which are "impossible" to be populated with samples in real world

LORE: Local Rule-Based Explanations of Black Box Decision Systems

Simplify the process of rule generation

Local MSE/Accuracy

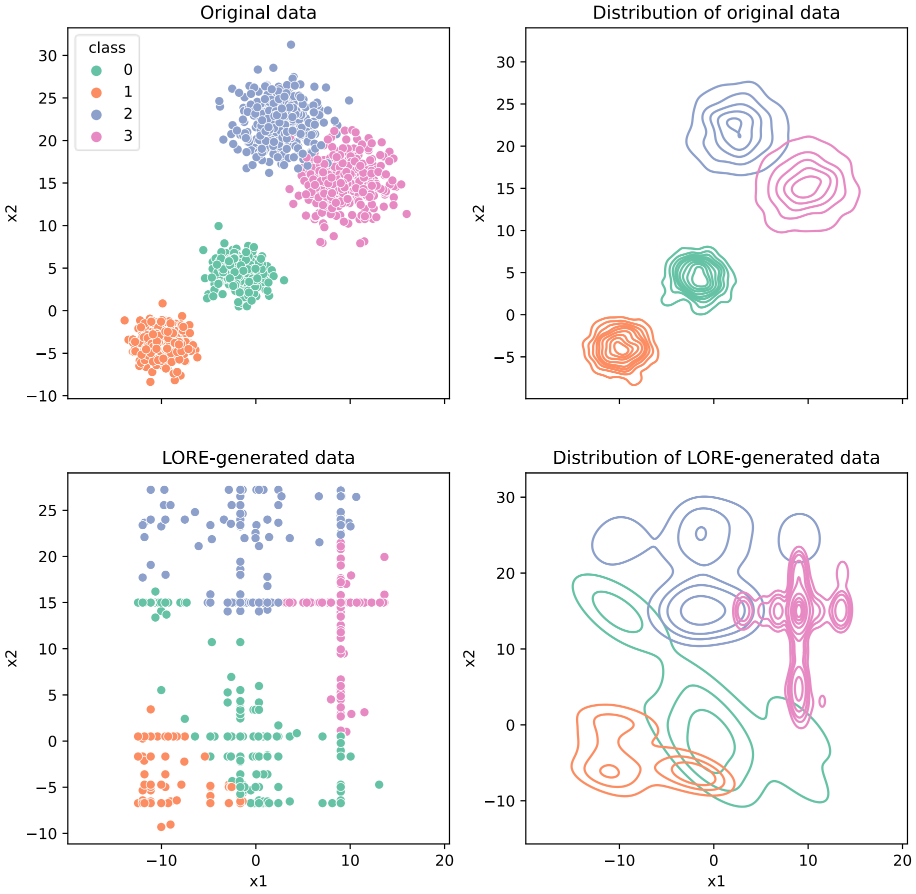
Balc-box model

Complexity limited by the depth of the decision tree

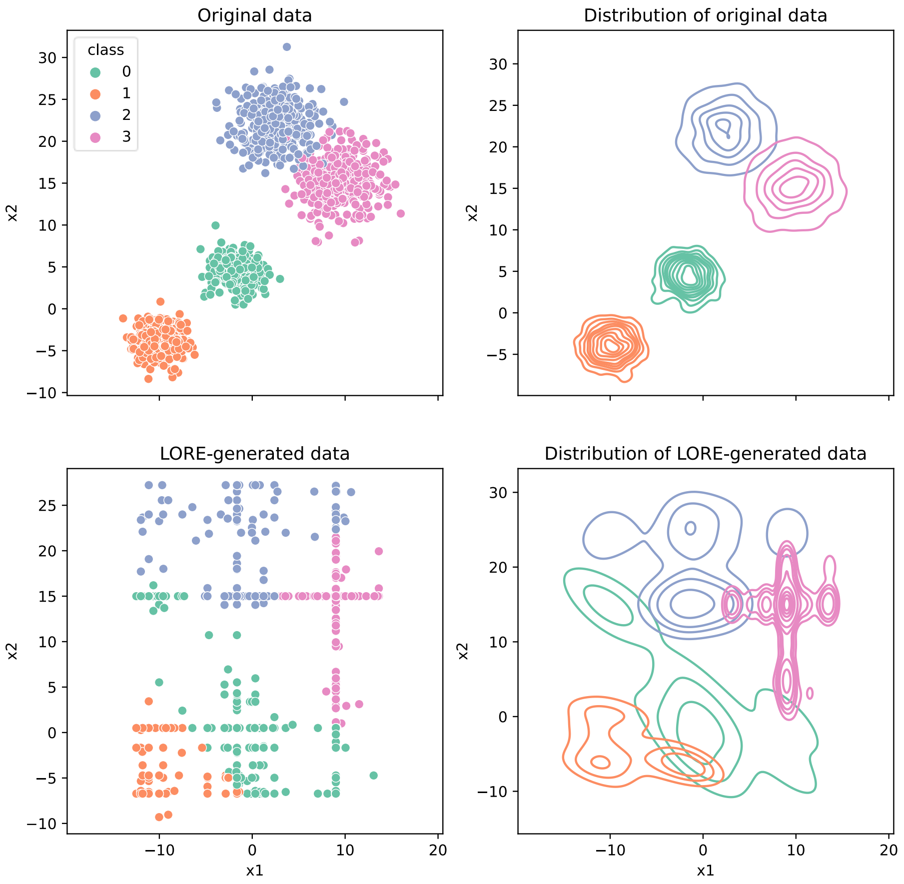
$$\phi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

Decision tree

Neighborhood: generated by genetic algorithm perturbaiton of original sample



Dataset generation and explanation creation

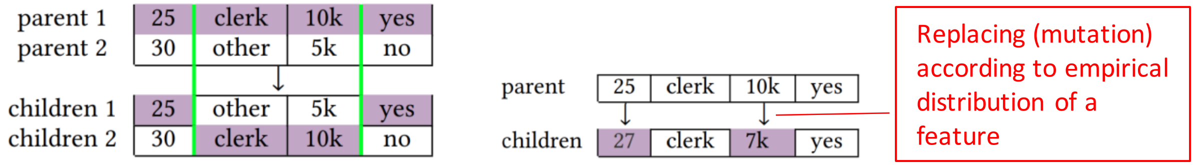


- Generate samples by optimizing following fitness functions:

$$fitness_{=}^{x}(z) = I_{b(x)=b(z)} + (1 - d(x,z)) - I_{x=z}$$

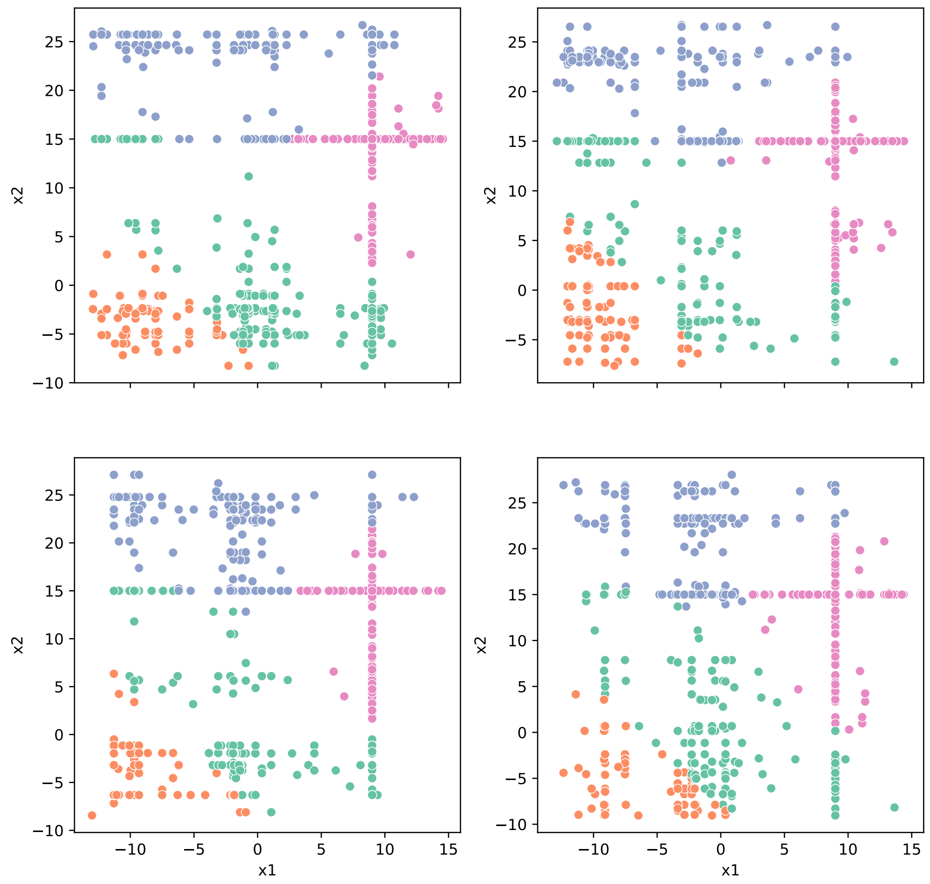
$$fitness_{\neq}^{x}(z) = I_{b(x)\neq b(z)} + (1 - d(x,z)) - I_{x=z}$$

- Fitness funcitons determine survivours and generaiton performs (2-point) crossover and mutation scheme:



- The resulting datset is balanced and complately artificial

Dataset generation and explanation creation



- Generate samples by optimizing following fitness functions:

$$fitness^{x}_{=}(z) = I_{b(x)=b(z)} + (1 - d(x,z)) - I_{x=z}$$

$$fitness^{x}_{\neq}(z) = I_{b(x)\neq b(z)} + (1 - d(x,z)) - I_{x=z}$$

- Fitness funcitons determine survivours and generaiton performs (2-point) crossover and mutation scheme:

parent 1	25	clerk	10k	yes
parent 2	30	other	5k	no
children 1	25	other	5k	yes
children 2	30	clerk	10k	no

parent	25	clerk	10k	yes
children	27	clerk	7k	yes

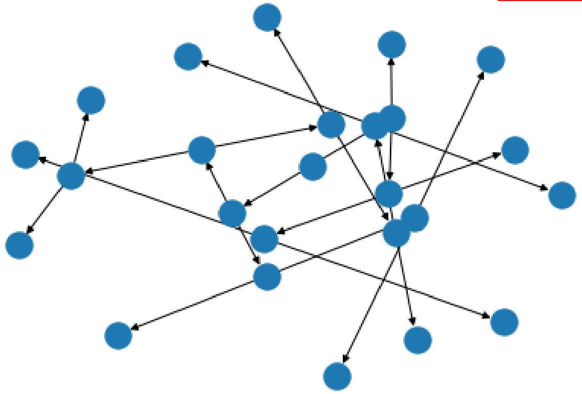
Replacing (mutation) according to empirical distribution of a feature

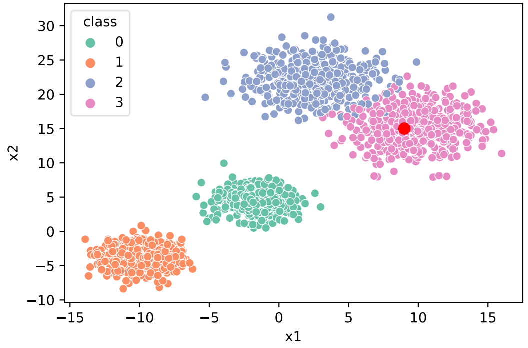
- The resulting datset is balanced and complately artificial

Dataset generation and explanation creation

```
([
  {'class': 3},
  { 'x1': '>6.532643',
    'x2': '-0.82784< x2 <=20.426761'},
  [499.0, 1.3]
 ],
 [
  {'x2': '<=-0.82784'}
 ])
```

Vizualization of the tree is rather cumbersome... :)



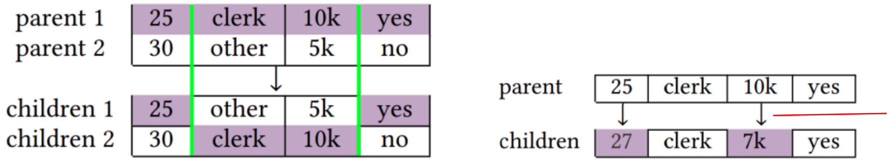


- Generate samples by optimizing following fitness functions:

$$fitness^{x}_{=}(z) = I_{b(x)=b(z)} + (1 - d(x,z)) - I_{x=z}$$

$$fitness^{x}_{\neq}(z) = I_{b(x)\neq b(z)} + (1 - d(x,z)) - I_{x=z}$$

- Fitness funcitons determine survivours and generaiton performs (2-point) crossover and mutation scheme:



Replacing (mutation) according to empirical distribution of a feature

- The resulting datset is balanced and complately artificial

Pros and cons

- Advantages
  - Fast and understandable implementation
  - Counterfactual generaiton (via traversing a decision tree which is surrogate model)
  - Rules generated with LORE have large coverage
- Disadvantages
  - Is based on data generation, hence can produce rule that are unintuitive in real world
  - Counterfactuals generated with LORE have low fidelity
  - Data generation may result in complately different explanations for the same instance

EXPLAN: Explaining Black-box Classifiers using Adaptive Neighborhood Generation

Simplify the process of rule generation

Local
MSE/Accuracy

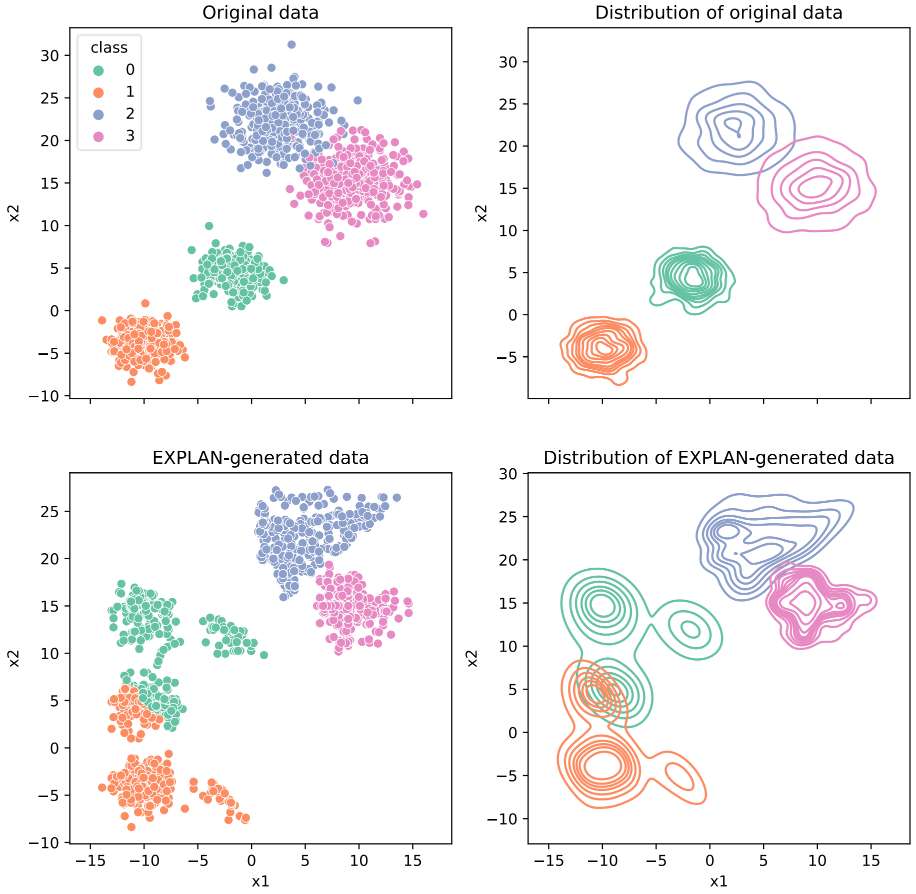
Balc-box model

Complexity limited by the depth of the decision tree

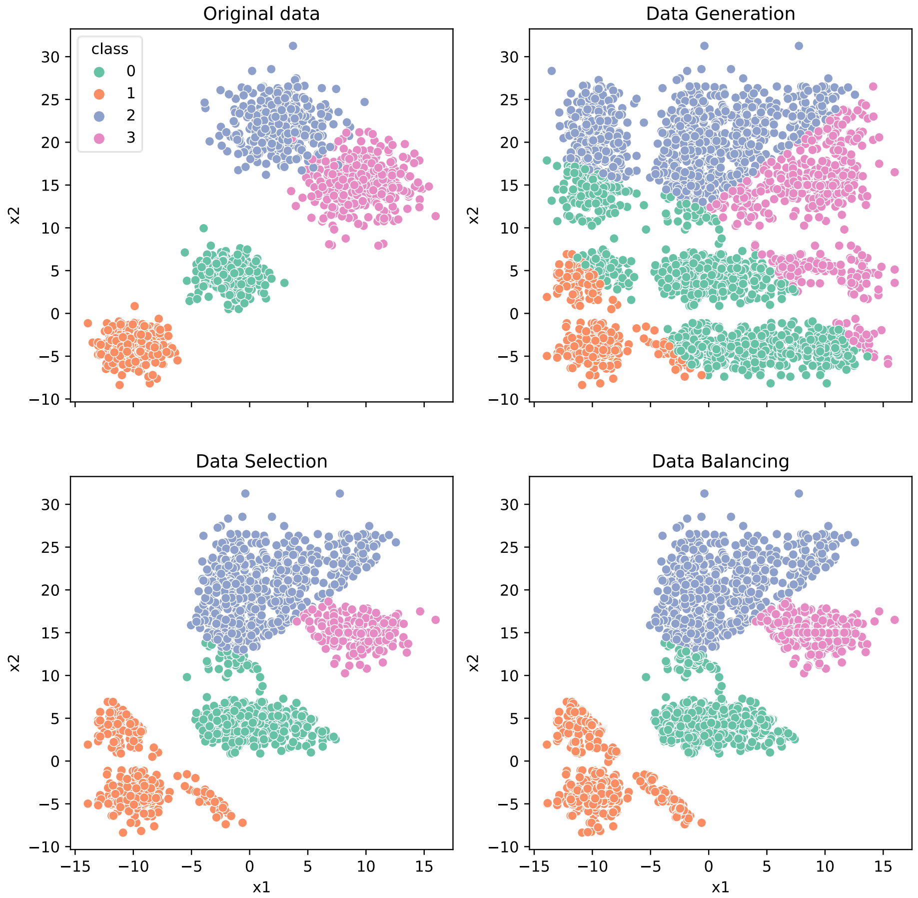
$$\phi(x) = \arg\min_{g\in G} L(f, g, \pi_x) + \Omega(g)$$

Decision tree

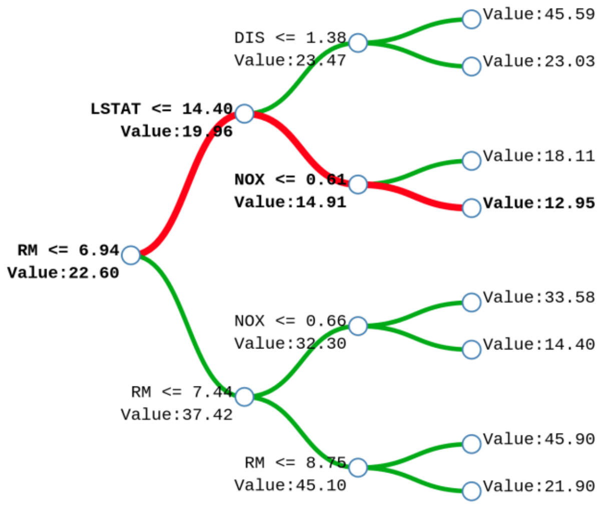
Neighborhood generated, manipulated and balanced



Data **generation**/manipulation/balancing

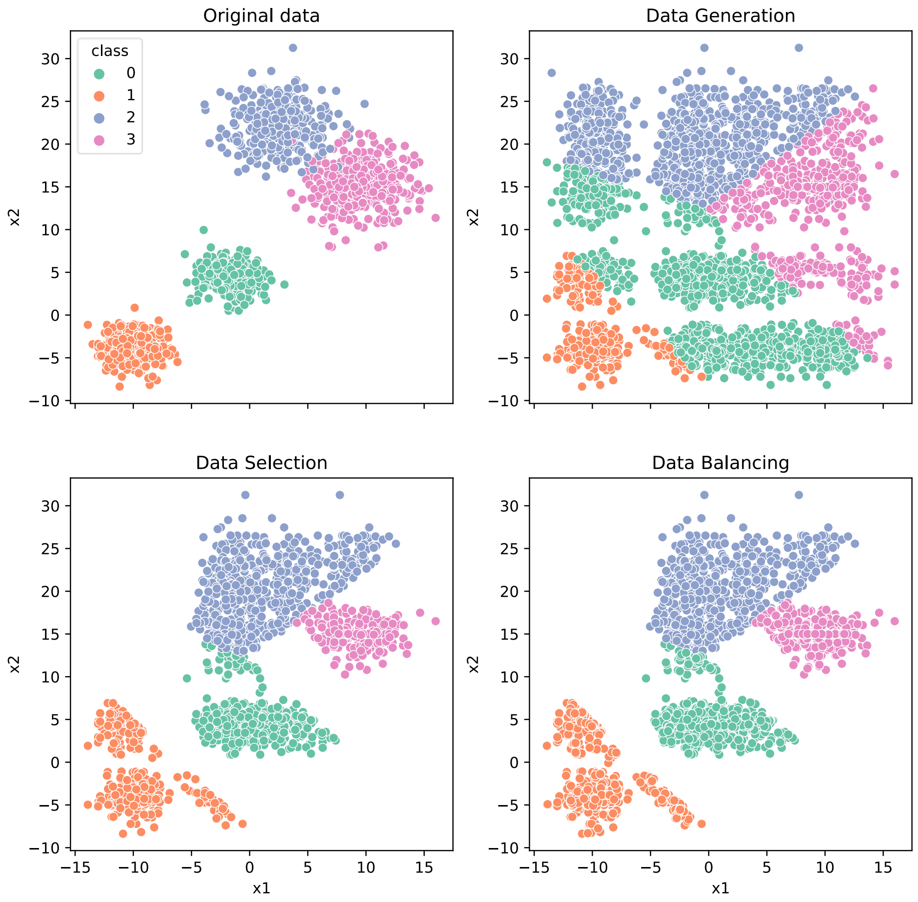


- Data is randomly sampled from the distribution of the original data
- RandomForest is trained on this data adn labels obtained from BlackBox
- Feature importances are obtained from random forest



Prediction: 21.90 ≈ 22.60 (trainset mean) + 14.82(gain from RM) + 7.68(gain from RM) - 23.2(loss from RM)

Data **generation**/selection/balancing

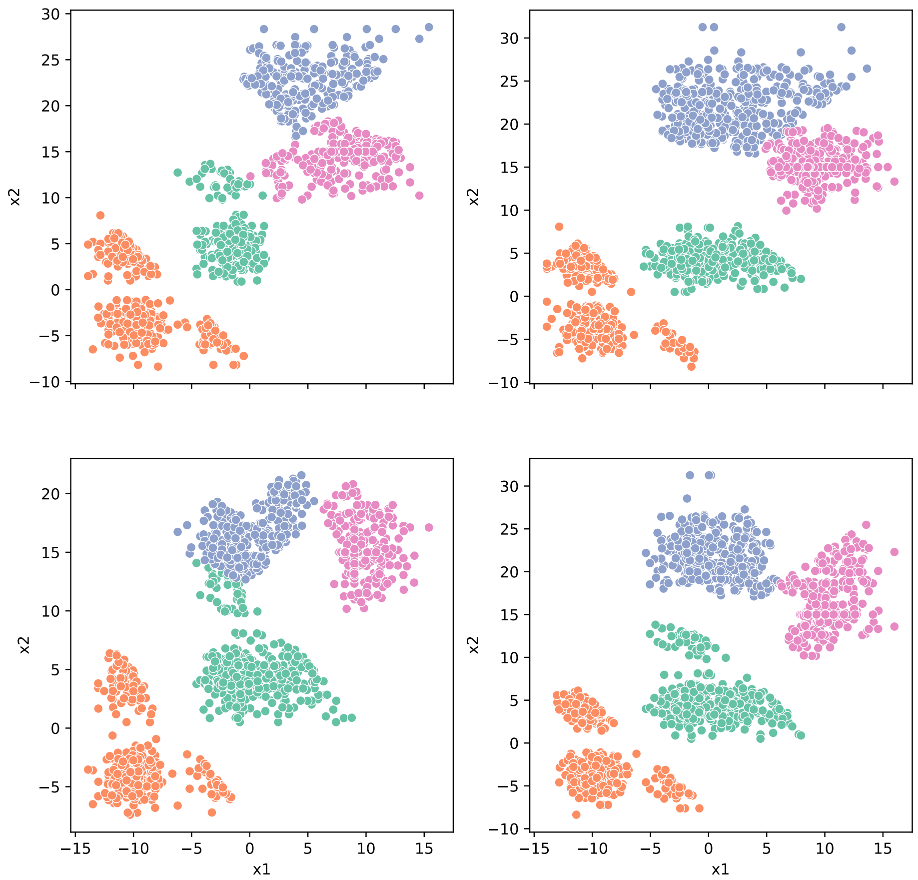


- Having importance (Γ) of instance *x* being explained and other samples *z*, we can refine dataset as follows (where j is a feature index):

$$z_j = \begin{cases} x_j & \text{if } \Gamma_{x_j} = \Gamma_{z_j} | x_j \neq z_j \\ z_j & \text{otherwise} \end{cases}$$

- The intuition behind that is that we want to make the genrated samples closer to the sample of interest and the distance is here not euclidean but more importance-based

Data **generation**/selection/balancing

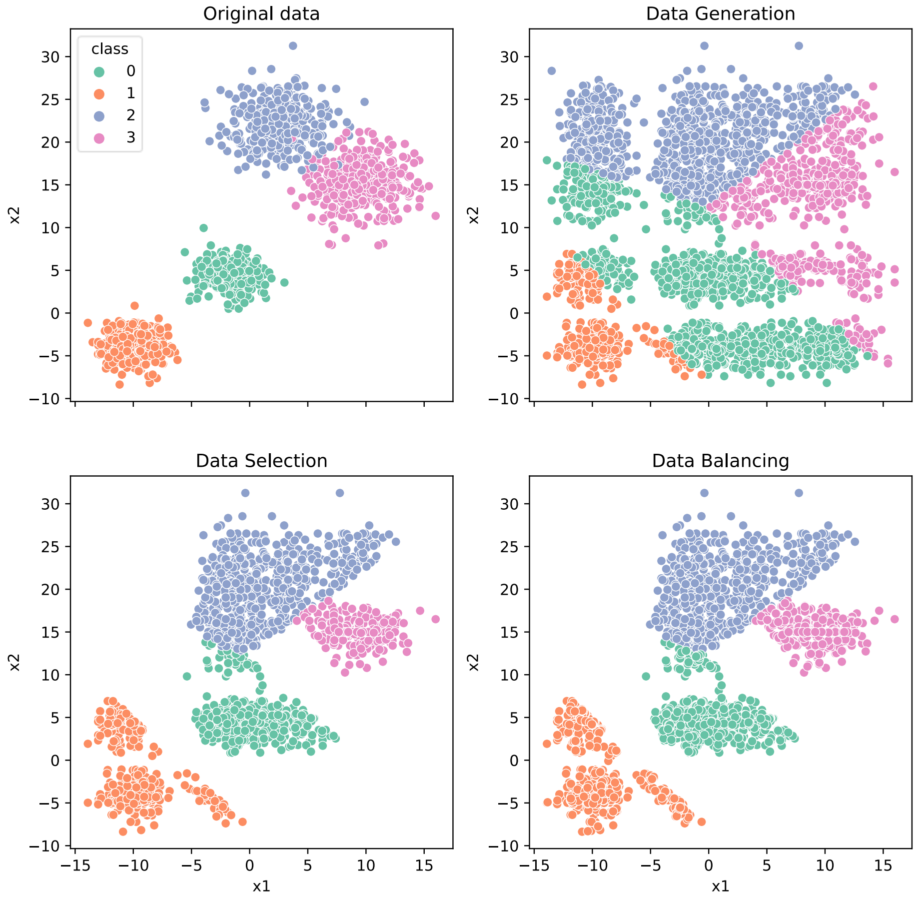


- Having importance (Γ) of instance *x* being explained and other samples *z*, we can refine dataset as follows (where j is a feature index):

$$z_j = \begin{cases} x_j & \text{if } \Gamma_{x_j} = \Gamma_{z_j} | x_j \neq z_j \\ z_j & \text{otherwise} \end{cases}$$

- The intuition behind that is that we want to make the genrated samples closer to the sample of interest and the distance is here not euclidean but
more importance-based
- Nontheless it is still stohastic process

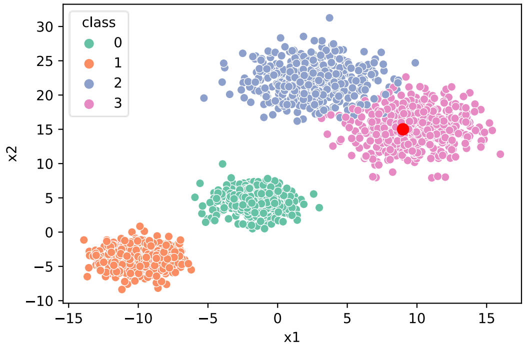
Data generation/**selection/balancing**

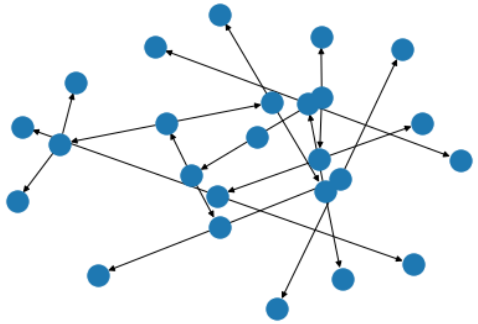


- Having the samples generated, the agglomerative clustering algorithm is used to select representative number of samples
- Clusterign is performed per class
- It is supposed to overcome the problem of KNN where appropriate K is not known
- Data balancing is perfomed with SMOTE

Explanation creation

```
[
  {'class': 3},
  {'x1': '>5.384601', 'x2': '<=18.633979'},
  [379.0, 1.3]
]
```





```
1: {'f_6': ['<=0.051216'], 'f_1': ['>-0.118408']}
2: {'f_6': ['<=0.034339'], 'f_3': ['<=0.231442'],
    'f_1': ['>-0.387023']}
3: {'f_6': ['<=-0.156712'], 'f_3': ['<=0.457593'],
    'f_1': ['>-0.619056']}
4: {'f_1': ['>-0.052240'], 'f_6': ['<=-0.101768']}
5: {'f_6': ['<=0.075657'], 'f_7': ['<=1.201282']}
```

EXPLAN

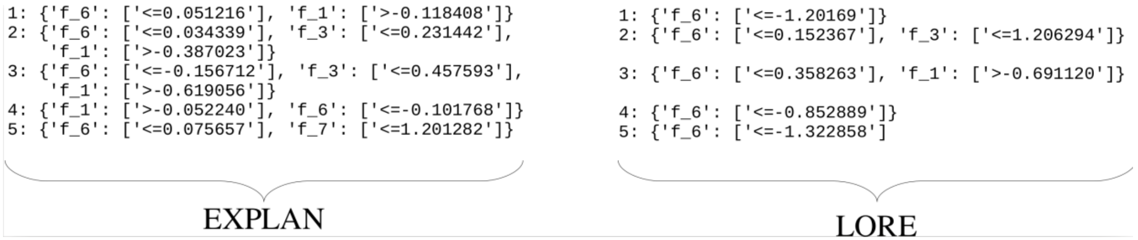
```
1: {'f_6': ['<=-1.20169']}
2: {'f_6': ['<=0.152367'], 'f_3': ['<=1.206294']}

3: {'f_6': ['<=0.358263'], 'f_1': ['>-0.691120']}

4: {'f_6': ['<=-0.852889']}
5: {'f_6': ['<=-1.322858']
```

LORE

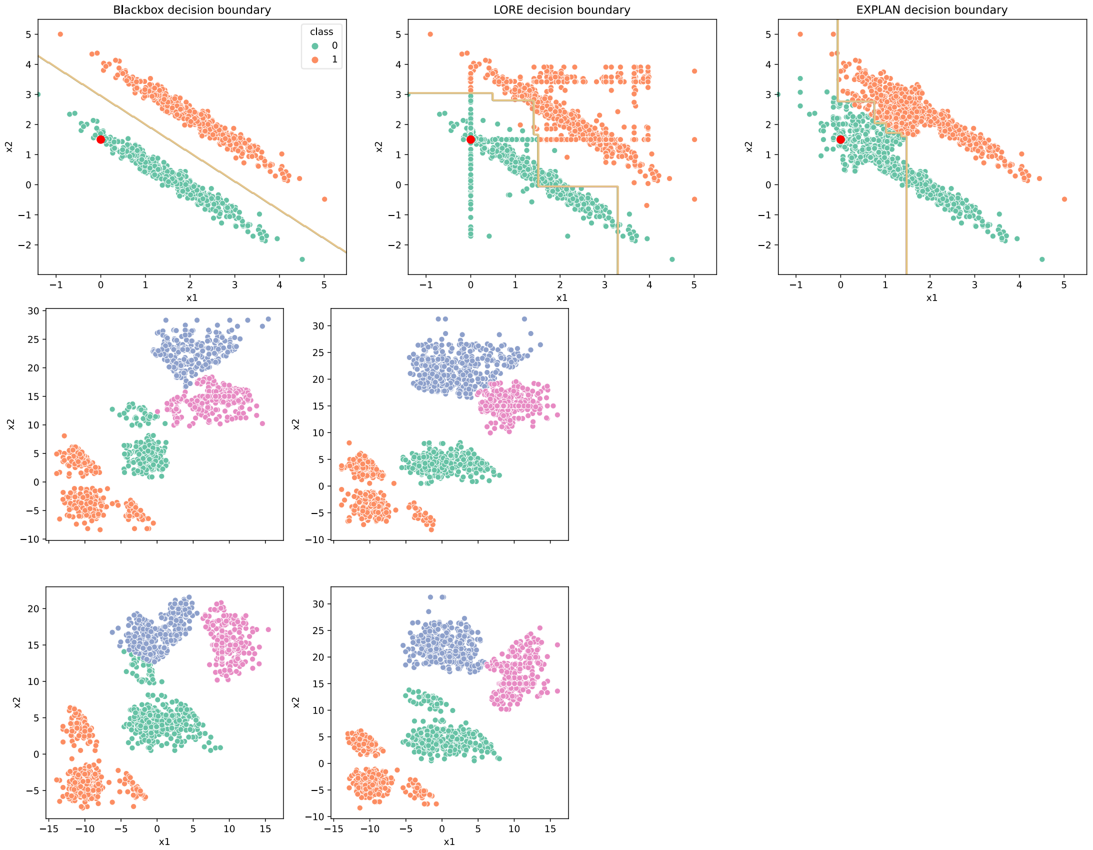
- The explanation is created using the same decision tree algorithm as LORE
- Therefore, the visualization of a tree is a bit bizare
- Supriosingly the implementaiton of EXPLAN do not provide counterfactuals, which could be extracted
- The generation process may result in different explantions for the same instance

Pros and Cons

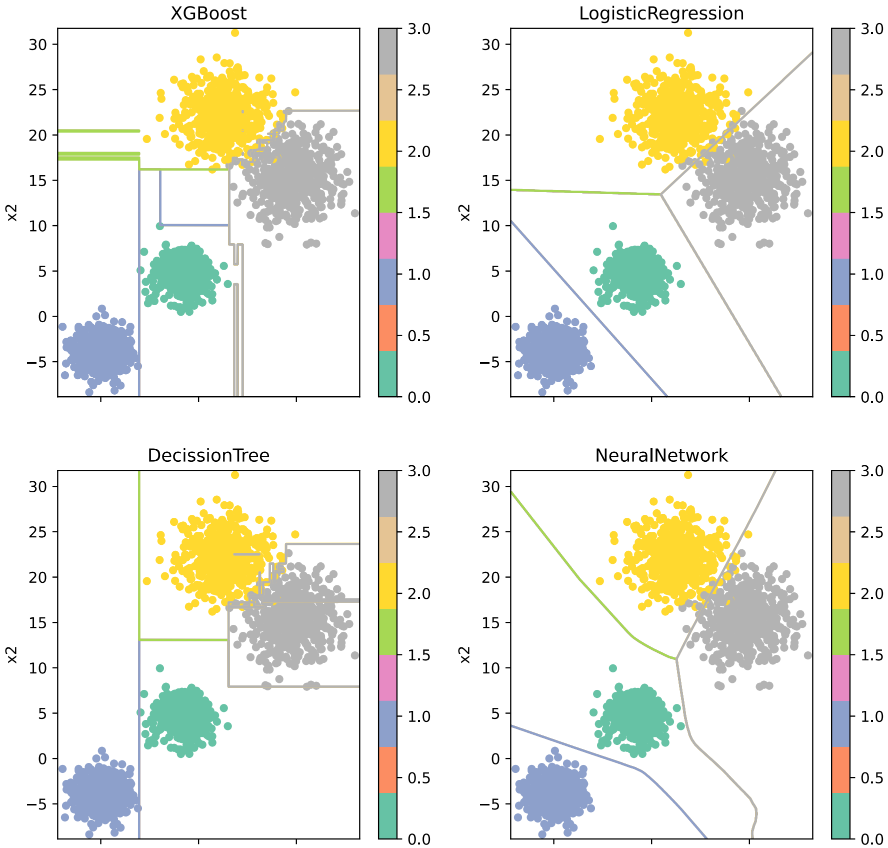
- Advantages
  - Same as LORE
  - Data generation makes it is a bit more stable in terms of providing similar explanations to similar instances



- Disadvantages
  - Same as LORE

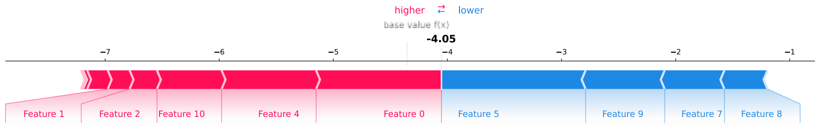


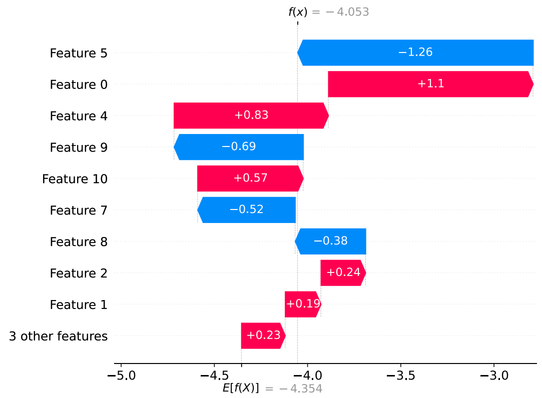
Rashomon effect

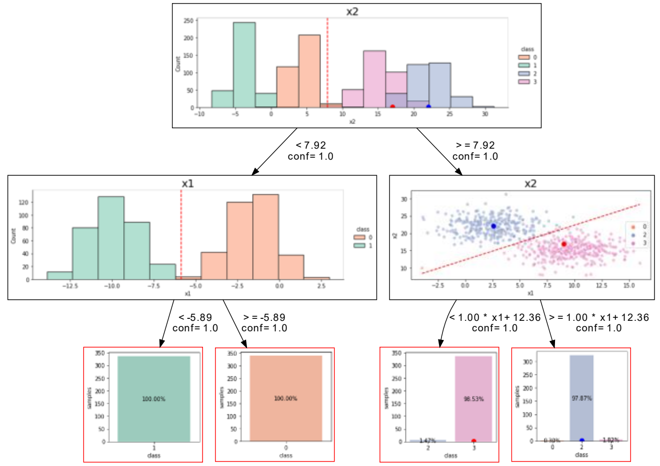


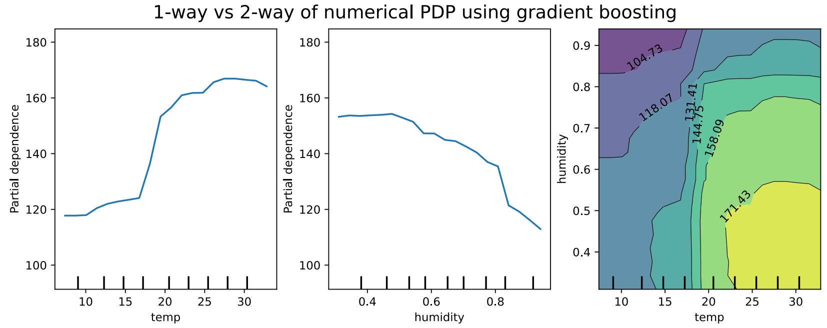
- Many models may be "right" but use very different methods to derive the "right"
- In Explainability we care about how the "right" is derived
- In such a case the more Rashomon effect the more doomed we are

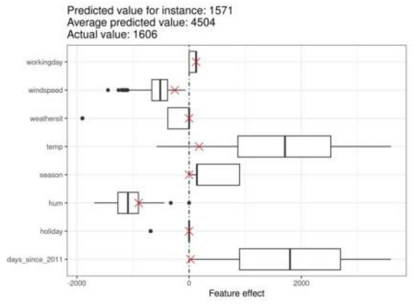
Multiple diffrent XAI methods

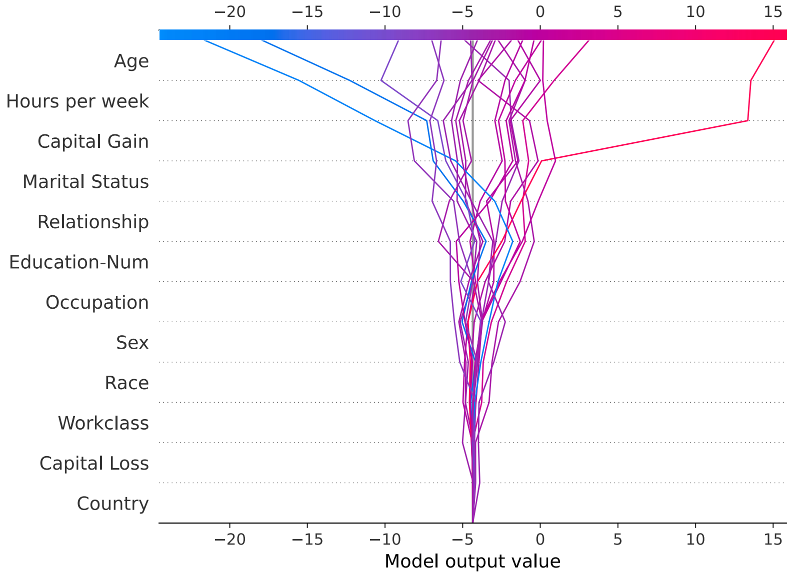


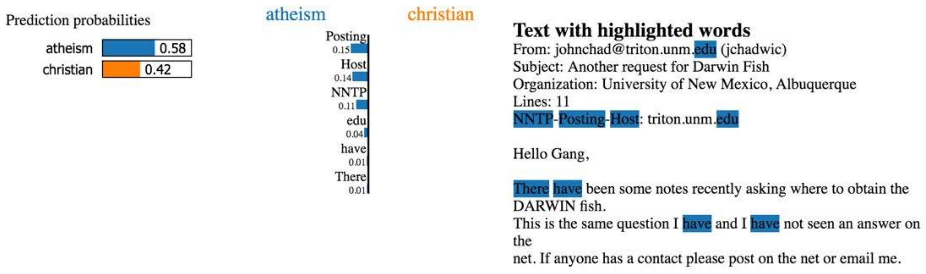


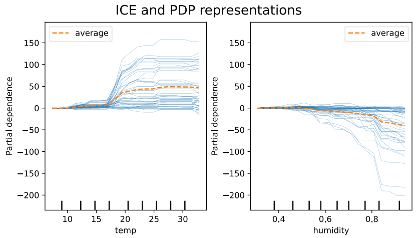




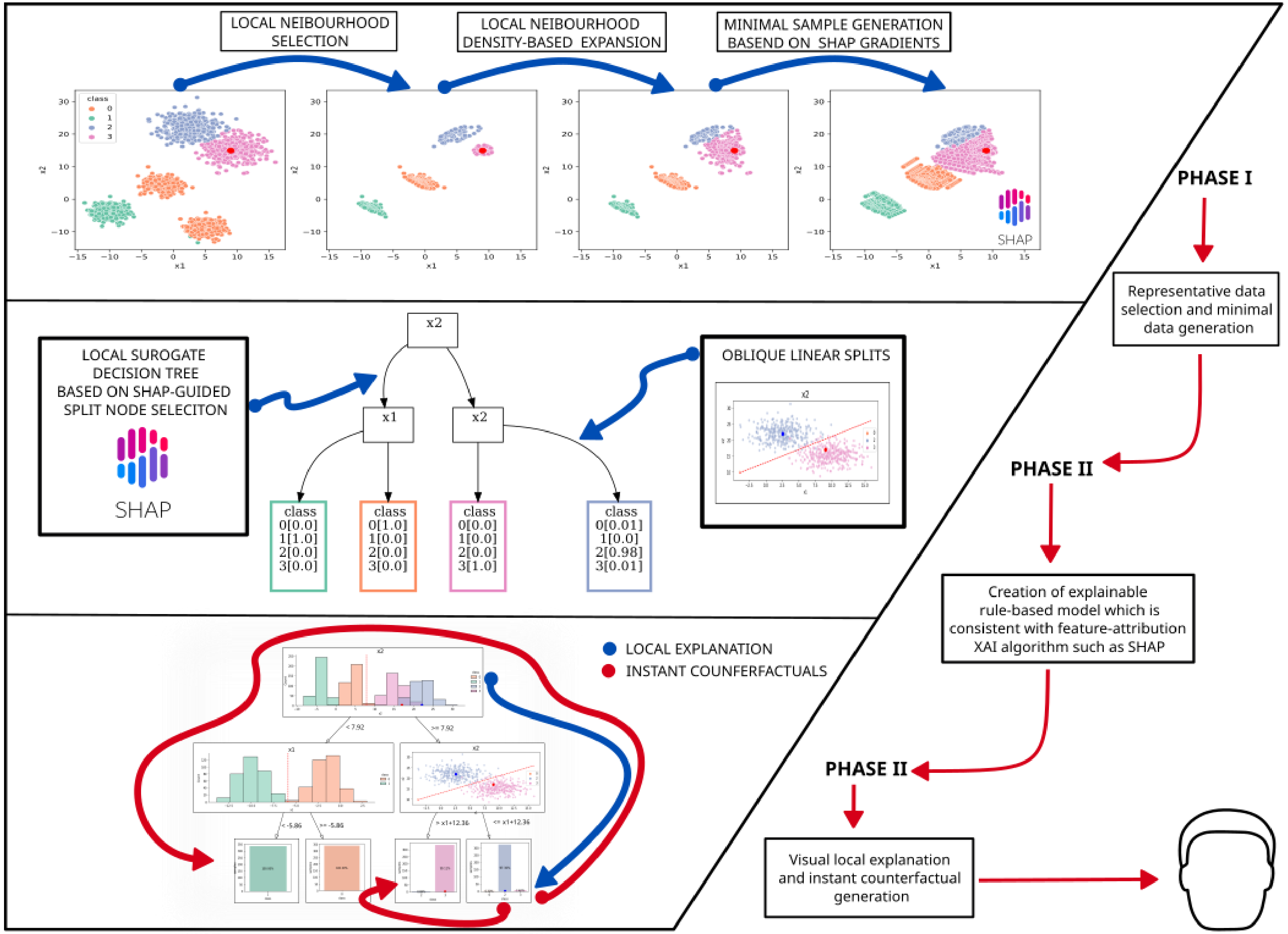






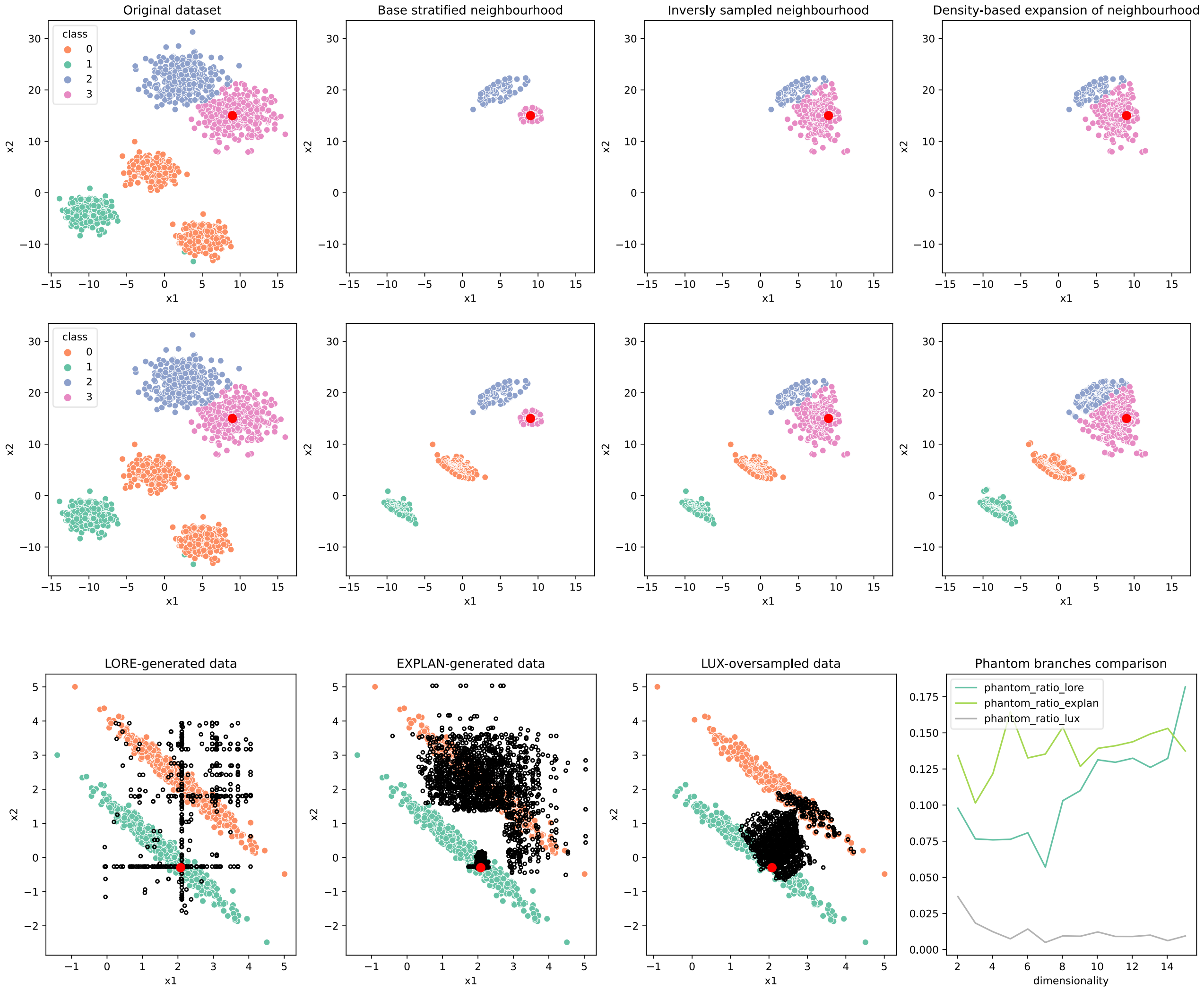


LUX: Local Universal Rule-based Explainer



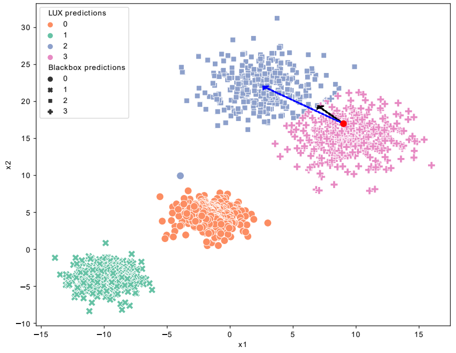
LOCAL NEIBOURHOOD SELECTION
LOCAL NEIBOURHOOD DENSITY-BASED EXPANSION
MINIMAL SAMPLE GENERATION BASEND ON SHAP GRADIENTS
SHAP
PHASE I
Representative data selection and minimal data generation
LOCAL SUROGATE DECISION TREE BASED ON SHAP-GUIDED SPLIT NODE SELECITON
SHAP
x2
x1
x2
class 0[0.0] 1[1.0] 2[0.0] 3[0.0]
class 0[1.0] 1[0.0] 2[0.0] 3[0.0]
class 0[0.0] 1[0.0] 2[0.0] 3[1.0]
class 0[0.01] 1[0.0] 2[0.98] 3[0.01]
OBLIQUE LINEAR SPLITS
PHASE II
Creation of explainable rule-based model which is consistent with feature-attribution XAI algorithm such as SHAP
LOCAL EXPLANATION
INSTANT COUNFERFACTUALS
< 7.92
>= 7.92
< -5.86
>= -5.86
> x1+12.36
<= x1+12.36
PHASE II
Visual local explanation and instant counterfactual generation

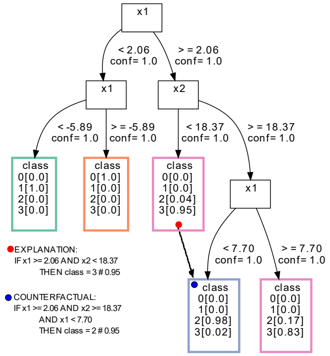
Neighbourhood generation and tree creation

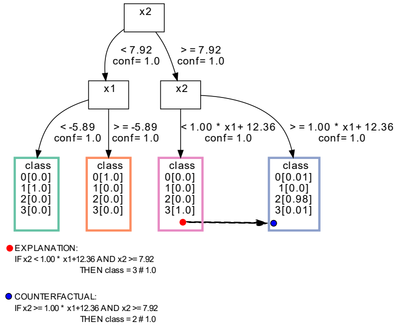


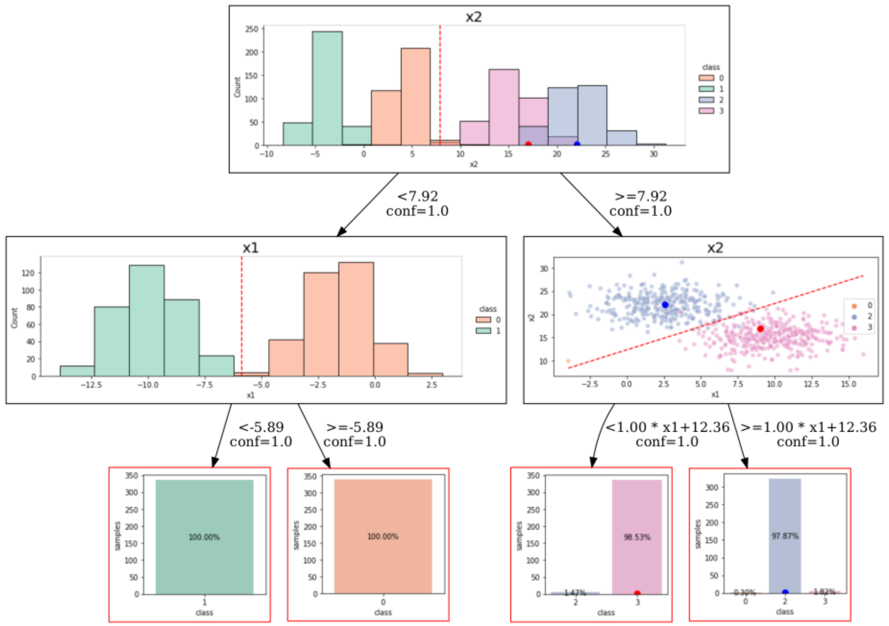
- Select Nearest Neighbours with KNN
- Expand NN by adding closest samples from oposite classess
- Expand NN by adding high density areas that "touches" already sampled data
- Generate samples around uncertain posints (possibly near decision boundaries) and in directions that point gradients of SHAP values
- Use these data to build a decision tree
- A tree uses information gain and SHAP-importances to select best splits

Explanation creation and vizualization



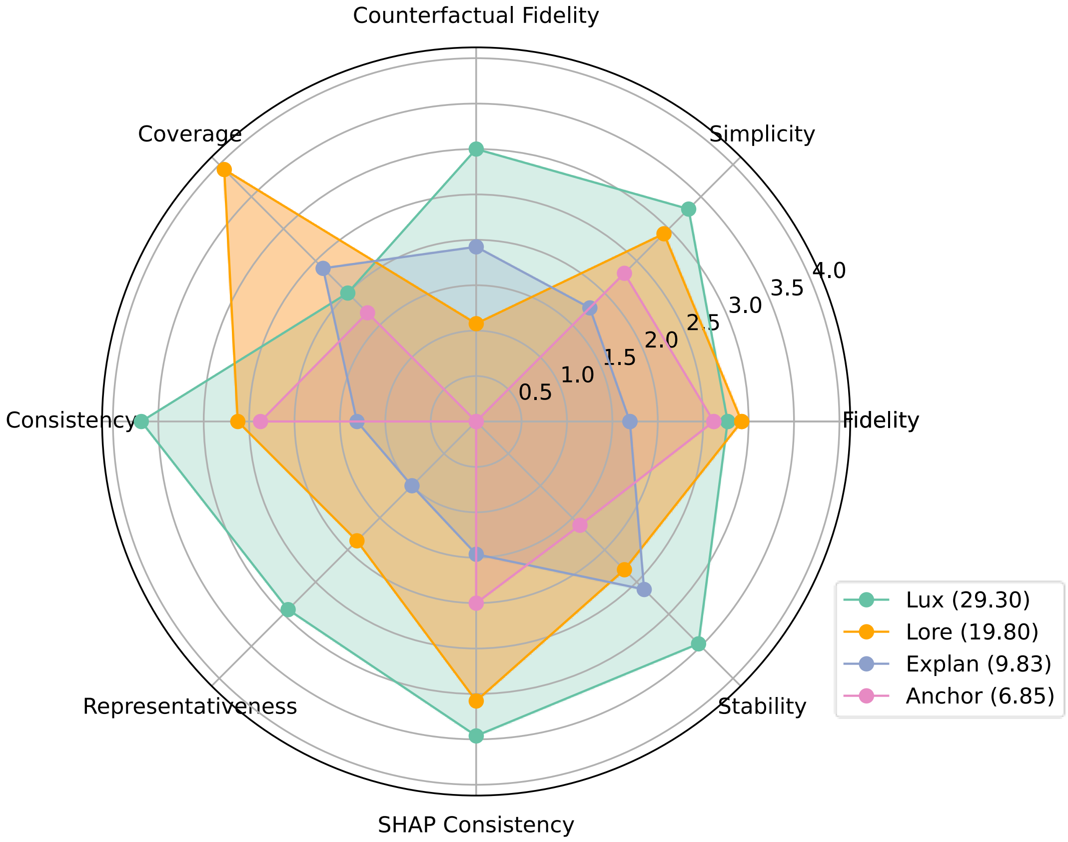






- Explanation is generated by extracting and pruning branch that the instace to explain falls into
- Counterfacutal is genrated by finding branch of opposite class that median/mean/nearest element is neares neighbour of instance to explain
- Oblique splits fit logistic regression using two most important features. This reduces depth

Comparison of all of them



Thank you for your attention!

JAGIELLONIAN UNIVERSITY
IN KRAKÓW

https://geist.re