Map of Explainability Approaches

Model types

- Logistic / Linear regression
- Decision Trees
- K-Nearest Neighbours
- Rule-based learners
- Generative Additive Models
- Bayesian Models

Transparent Models

Explainability Approaches

Opaque Models

- Random Forest
- Support Vector Machines
- Multi-layer Neural Network

Post-Hoc Explainability

Explainability Categories

Model-Agnostic

Model-Specific

- Explanation by Simplification
- Feature relevance explanation
- Local explanations
- Visual explanations
- Explanation by Simplification
- Feature relevance explanation

Explainability Principles

- Rule-based learner
- Decision tree
- Influence functions
- Sensitivity
- Game theory inspired
- Interaction based
- Rule-based learner
- Linear approximation
- Counterfactuals
- Sensitivity
- Dependency plots
- Rule-based learner
- Decision trees / prototypes
- Distillation
- Feature importance

Popular Techniques (examples)

- SHAP
- Anchors
- LIME
- Counterfactual instances
- ICE
- PDP
- InTrees