ONE PAGE SUMMARY

IT & Software › Machine Learning › Explainability

Introduction to Explainability

Understanding Explainability in Machine Learning Models

DEFINITION

> Explainability refers to the methods and techniques used to make machine learning models understandable to humans. It helps users grasp how decisions are made by algorithms.

KEY CONCEPTS

Transparency

Transparency involves making the inner workings of a model clear and accessible to users.

Interpretability

Interpretability is the degree to which a human can understand the cause of a decision made by a model.

Feature Importance

Feature importance identifies which input variables most significantly influence the model's predictions.

Local vs Global Explainability

Local explainability focuses on individual predictions, while global explainability looks at the model's overall behavior.

Model-Agnostic Methods

These methods can be applied to any model type, providing insights regardless of the underlying algorithm.

EXAMPLES

→ Using LIME to explain a specific prediction of a classifier.

→ Visualizing decision trees to show how decisions are made.

→ Employing SHAP values to assess feature contributions in a model.

MEMORY TIPS

★ Remember 'TIF' for Transparency, Interpretability, and Feature Importance.

★ Think of 'Local' as 'Look at one' and 'Global' as 'Gaze at all'.

★ Use the acronym 'MAG' for Model-Agnostic methods.

COMMON MISTAKES

✗ Confusing interpretability with accuracy; they are not the same.

✗ Neglecting to consider the audience's technical background when explaining models.

✗ Overlooking the importance of visual aids in enhancing understanding.

QUICK RECAP

Explainability in machine learning is crucial for trust and understanding. It involves making models transparent and interpretable, focusing on both individual predictions and overall behavior. Key methods include feature importance and model-agnostic techniques.