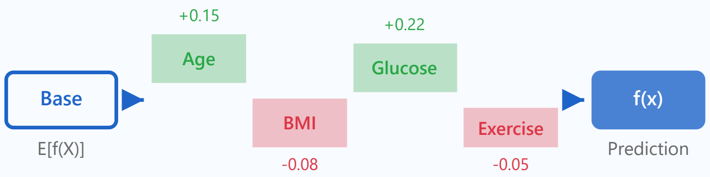
SHAP Values for Model Interpretation

SHapley Additive exPlanations - unified framework for interpretability

How SHAP Works



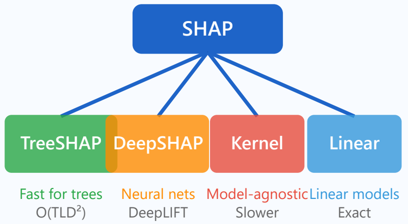
$$f(x) = \varphi_0 + \varphi_1 + \varphi_2 + \ldots + \varphi_n$$

Shapley Value: Fair contribution from cooperative game theory

Additive feature attribution

- Considers all possible feature combinations
- Satisfies consistency & local accuracy

SHAP Algorithms



Key Properties: ✓ Local accuracy ✓ Missingness ✓ Consistency

The only method satisfying all desired properties (Lundberg & Lee, 2017)

Visualization Types

Waterfall Plot

Single prediction

Summary Plot

Global importance

Dependence Plot

Feature effects

Force Plot

Interactive push/pull