Machine Learning Hierarchy

Level 1: The Fundamental Unit

Decision Tree

- **Concept:** A single flowchart structure representing sequential decisions.
- **Characteristics:** Weak and unstable on its own; prone to overfitting.
- **Role:** The fundamental "brick" used to construct robust ensemble models.

Level 2: The Strategy (Ensemble Learning)

Goal: Combine multiple "bricks" (trees) to create a strong structure.

Strategy A: Bagging (Bootstrap Aggregating)

- **Logic: Parallel** execution.
- **Mechanism:** Trains N trees independently. Each tree analyzes a random subset of the data. The final result is an average or vote.
- **Key Algorithm:**
  - **Random Forest:** Applies Bagging combined with Random Feature Selection.

Strategy B: Boosting

- **Logic: Sequential** execution.
- **Mechanism:** Iterative correction. Tree N is built specifically to fix the errors of Tree $N-1$.
- **Key Algorithms:**
  - **AdaBoost:** Adjusts **sample weights** (focuses on hard-to-classify data points).
  - **Gradient Boosting:** Uses **gradient descent** to minimize error residuals (focuses on reducing loss).

Level 3: Software Implementations (The Packages)

Software libraries that implement the Gradient Boosting algorithm.

Package	Key Characteristics
XGBoost	Optimized for performance; handles missing values (`NaN`) natively; widely used in science.
LightGBM	Histogram-based; uses leaf-wise growth; optimized for speed and large datasets.
HistGradientBoosting	Scikit-learn's native implementation of the histogram-based algorithm (similar to LightGBM).
CatBoost	Specialized for handling categorical data (text labels) directly without preprocessing.