REGRESSION I
KNN, DECISION TREE AND RANDOM FOREST

Dr. Lablanche Pierre-Yves
African Institute for Mathematical Sciences

REGRESSION vs. CLASSIFICATION

Two different problems… fundamentally not very different.

- **Classification problem :** predict category (discrete value)
- **Regression problem :** predict (continuous) value

Many classification algorithms have a regression version
(such as k-Nearest Neighbour, Decision Tree and Random Forest)

REGRESSION vs. CLASSIFICATION

Two different problems… fundamentally not very different.

- **Classification problem :** the target/label is categorical
- **Regression problem :** the target/label is continuous

Examples of continuous variables : Temperature, Market Value, etc.

KNN REGRESSION

For **classification** kNN performs a **majority vote** upon the *k* closest neighbours.

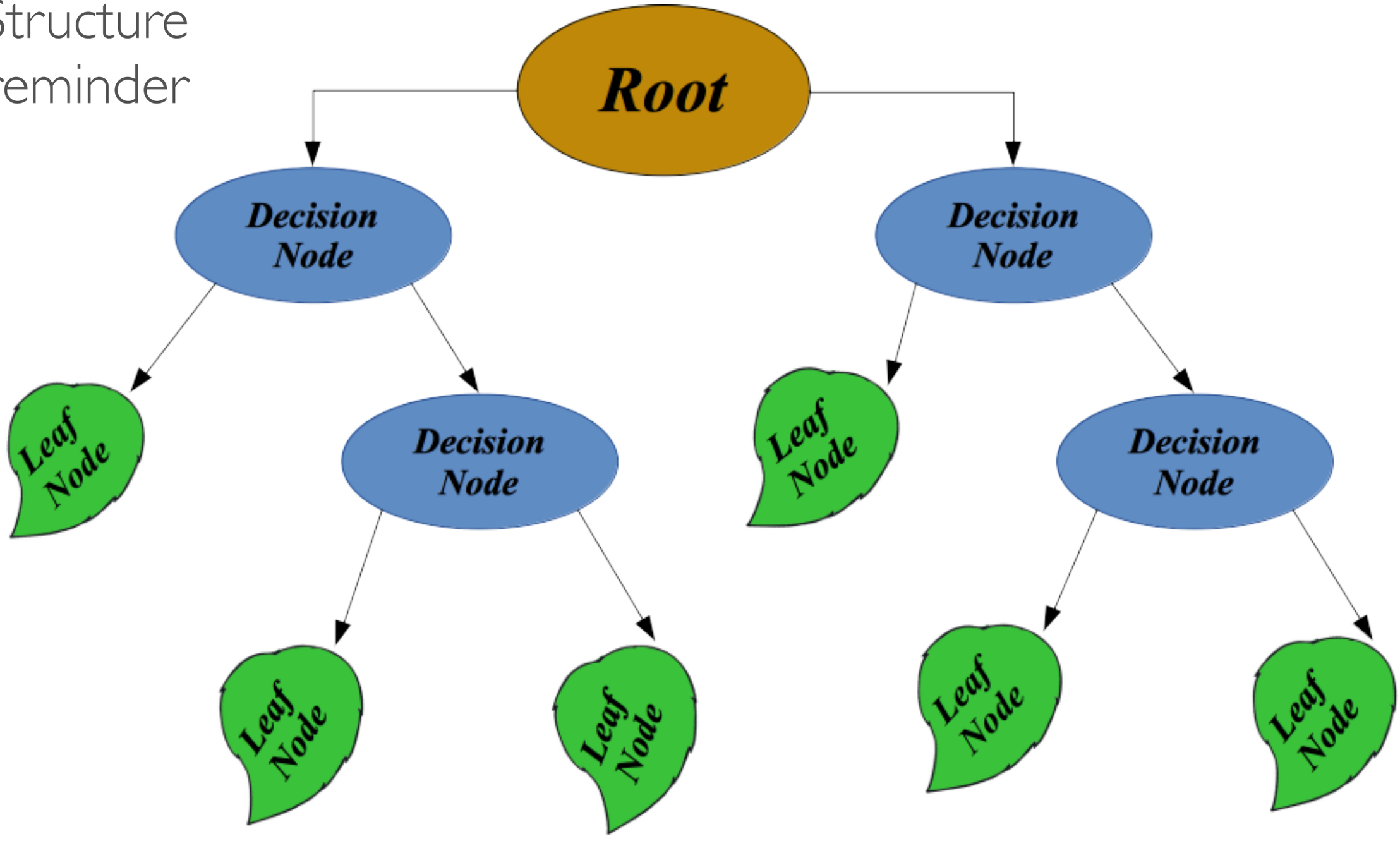
For **regression** kNN will compute the **average** (weighted or not) upon the *k* closest neighbours :

$$y = \frac{1}{k} \sum_{i=0}^{k} w_i \cdot y_i$$

Where the weights w_k can reflect the distance or the importance of different instances (or any other attribute of your choice).

DECISION TREE REGRESSION

Structure reminder



DECISION TREE REGRESSION

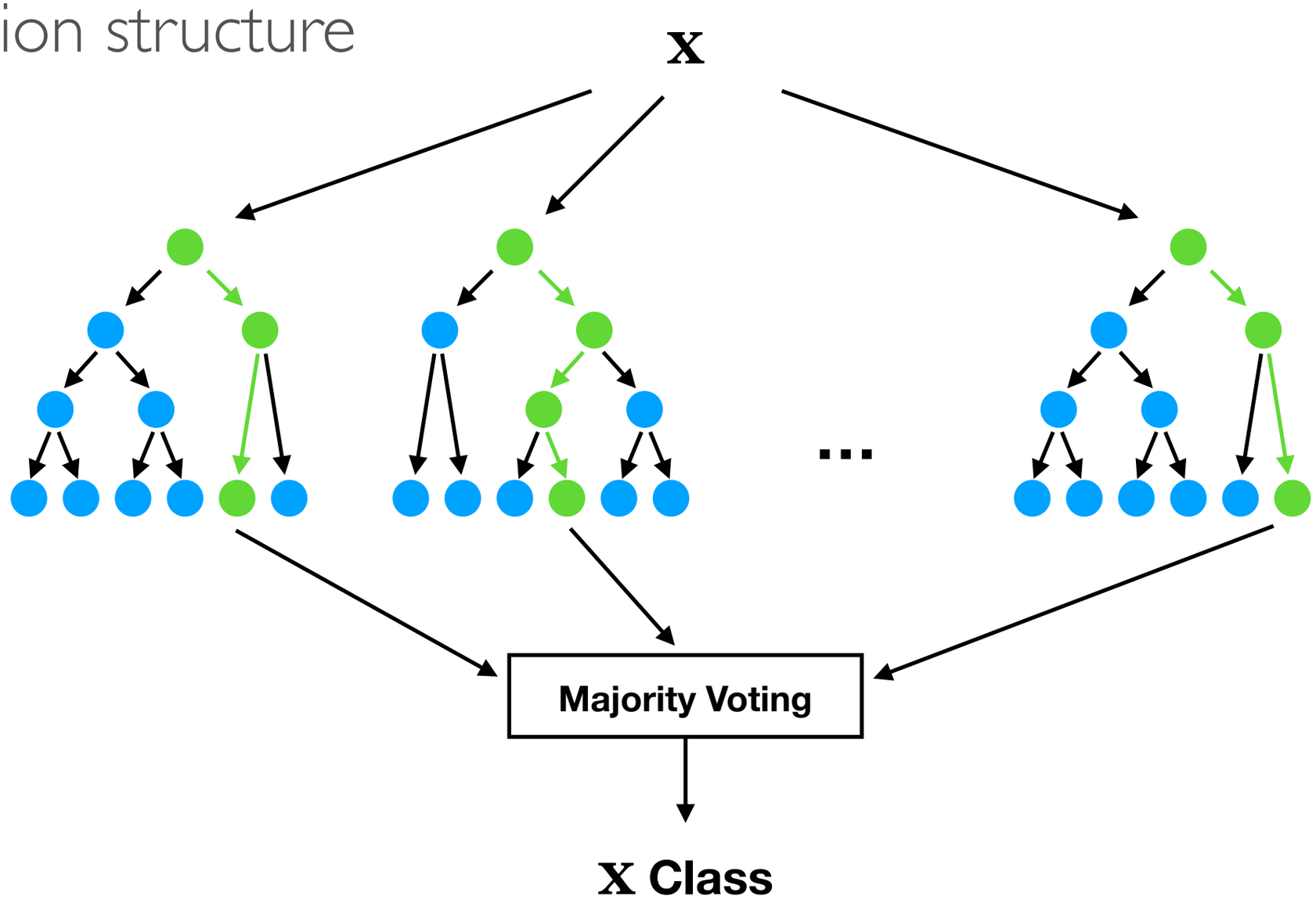
At a *leaf node*, instead of a **class**, a **continuous value** will be found.

The number of possible values will depend on the number of leaf nodes (discretisation).

Suffers the same drawbacks as the classification decision tree.

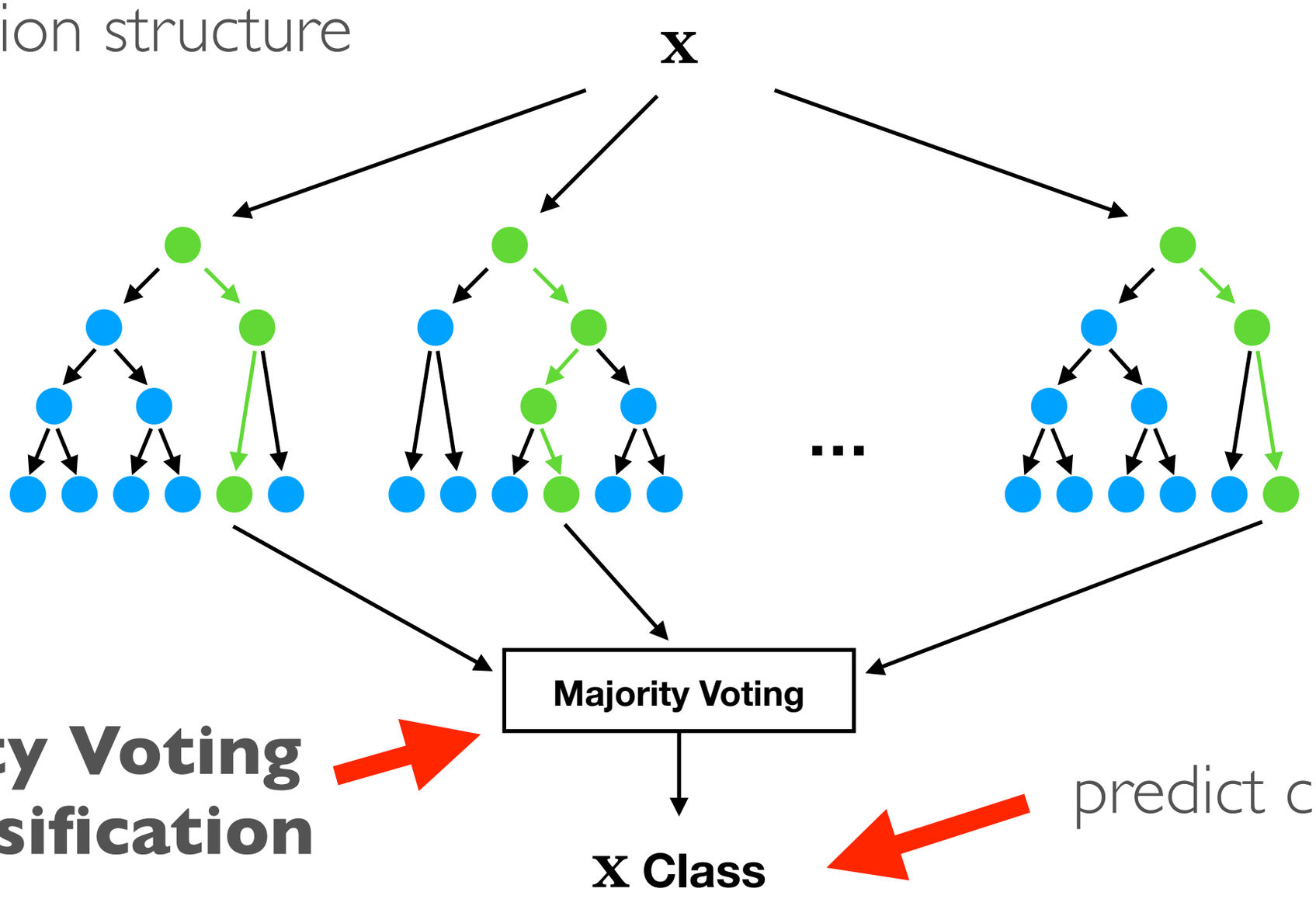
RANDOM FOREST REGRESSION

Recall the random forest classification structure



RANDOM FOREST REGRESSION

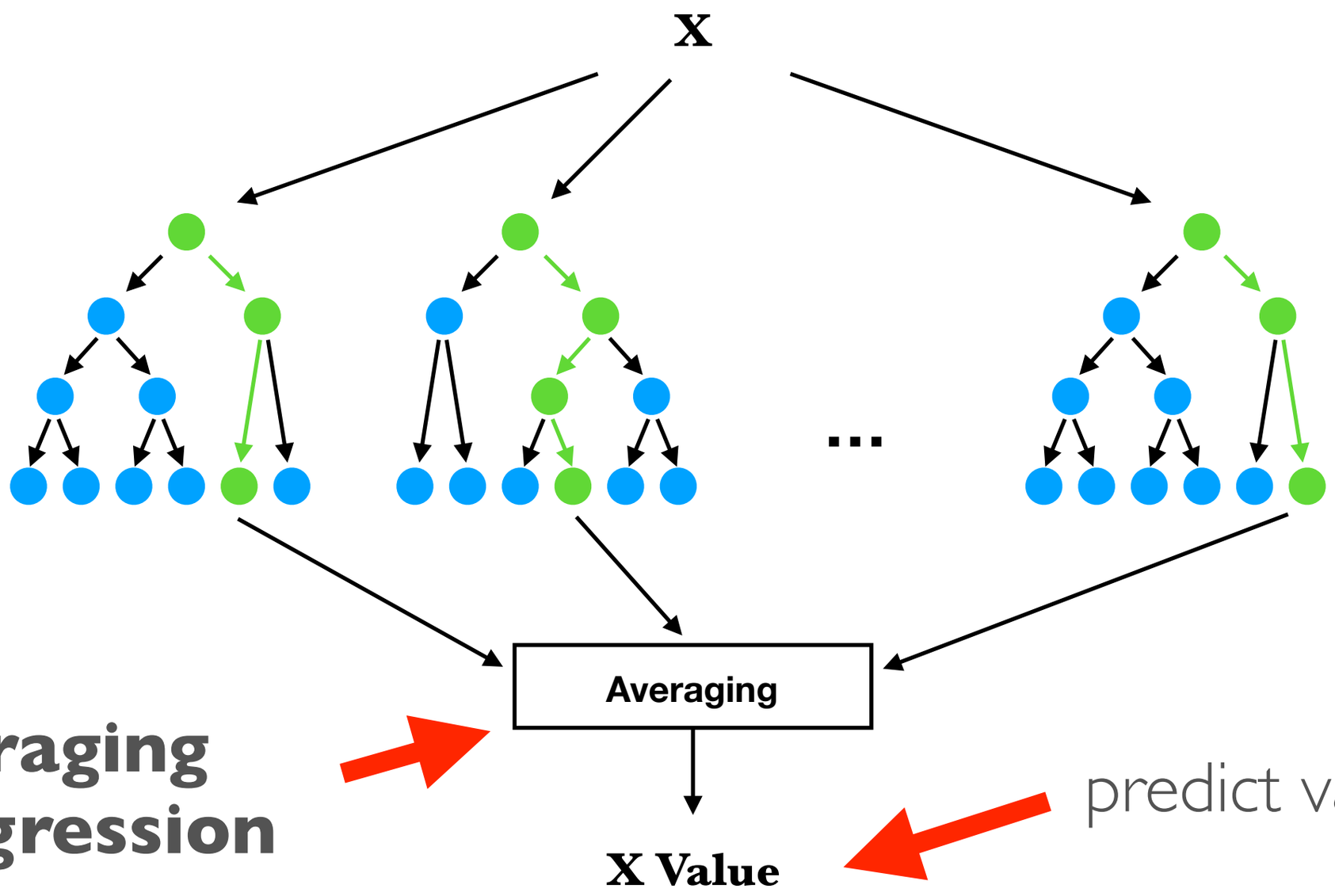
Recall the random forest classification structure



Majority Voting for classification

predict class

RANDOM FOREST REGRESSION



Averaging for regression

predict value

CODING TIME (1)

- Load the **`boston`** dataset available in scikit-learn.
- Use the **`KNeighborsRegressor`**, **`DecisionTreeRegressor`** and **`RandomForestRegressor`** algorithms to predict test values
- Compare Results