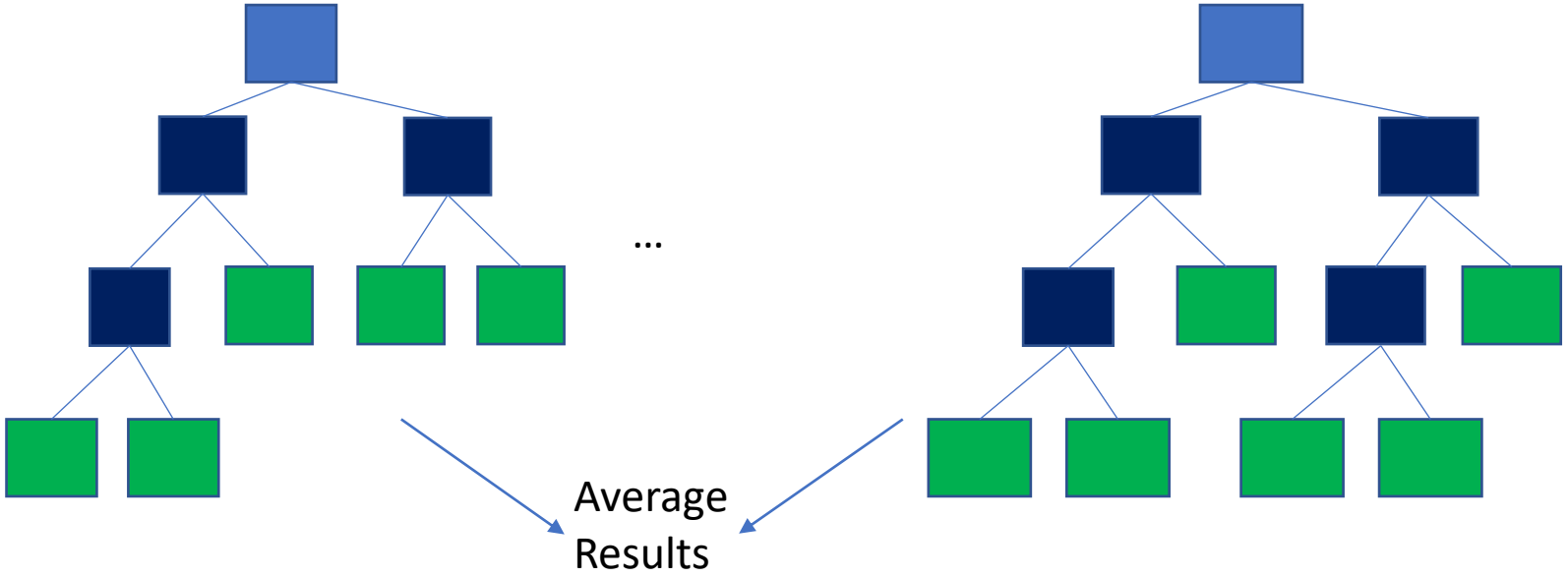
Random Forests 101

Random Forests

Introduction

- Based on decision trees
- Creates many trees instead of just one
- Averages the results



Decision Trees

Hyperparameter

Number of Trees

- Defines how many trees should be created
- Generally: tree number up → model performance up, but higher computational effort

Number of Variables mtry

- Number of variables sampled as candidates at each split

Random Forests

Advantages / Disadvantages

\+

- Single trees often underfit or overfit

→ RF finds good balance

- Cover non-linearity
- Stable results (e.g. low impact of outliers)
- High performance
- Provides feature importance

\-

- Black-box
- High computational effort
- Bias for variables with many categories