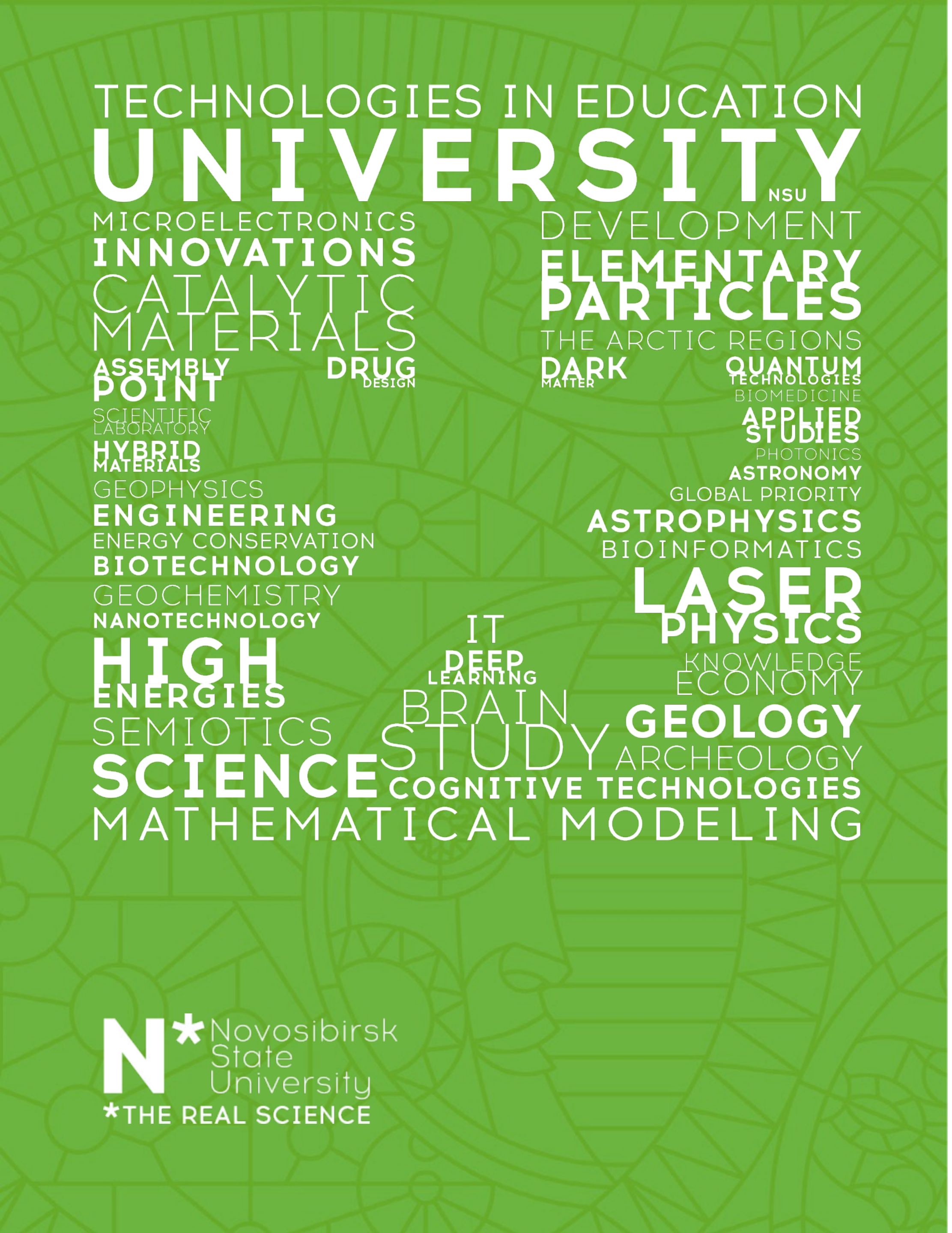


Decision Tree & Random Forest

Luu Minh Sao Khue
Department of Mathematics and Mechanics

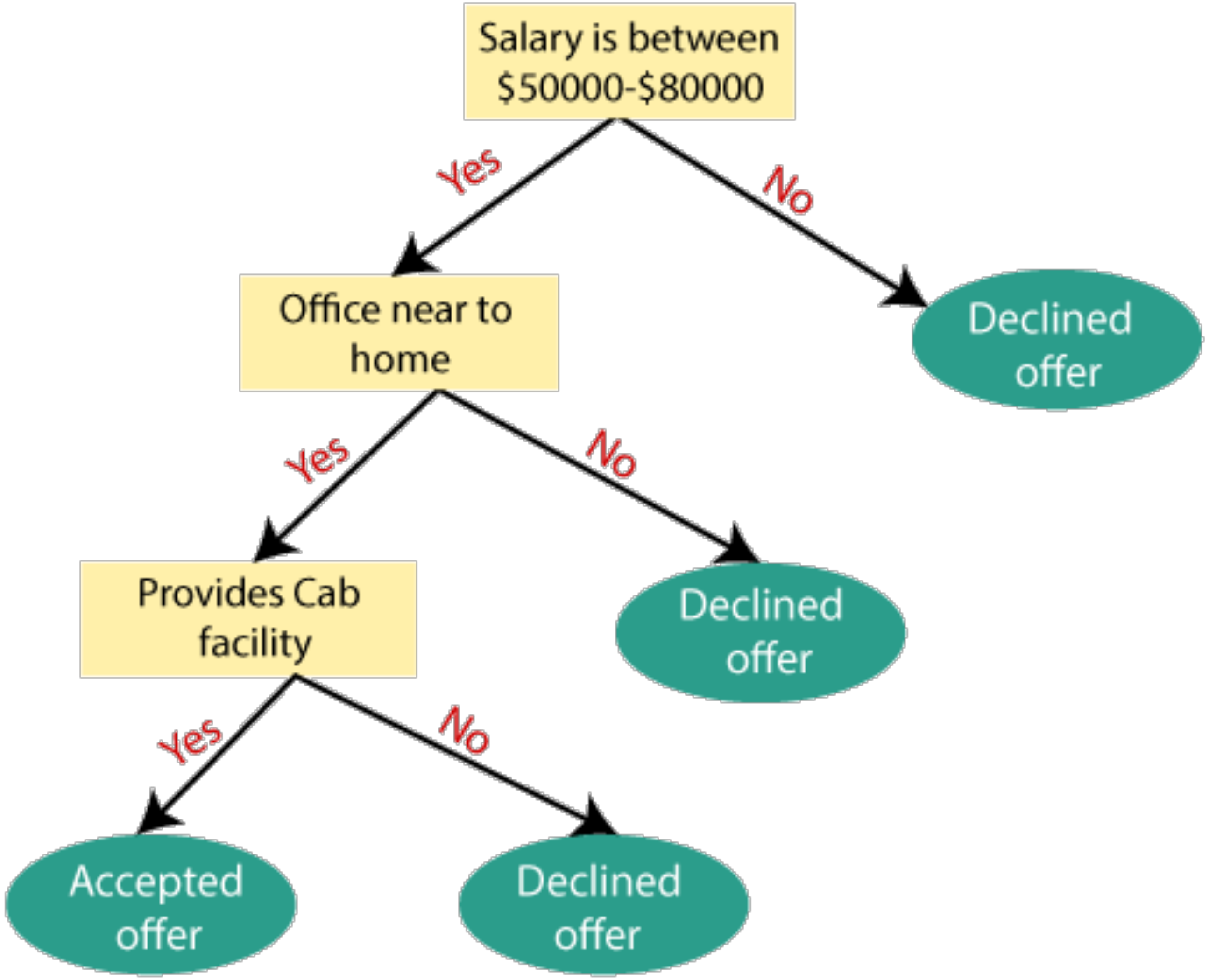
Overview

- Decision Tree
- Ensemble Strategy: Bagging
- Random Forest

Decision Tree

Decision Tree

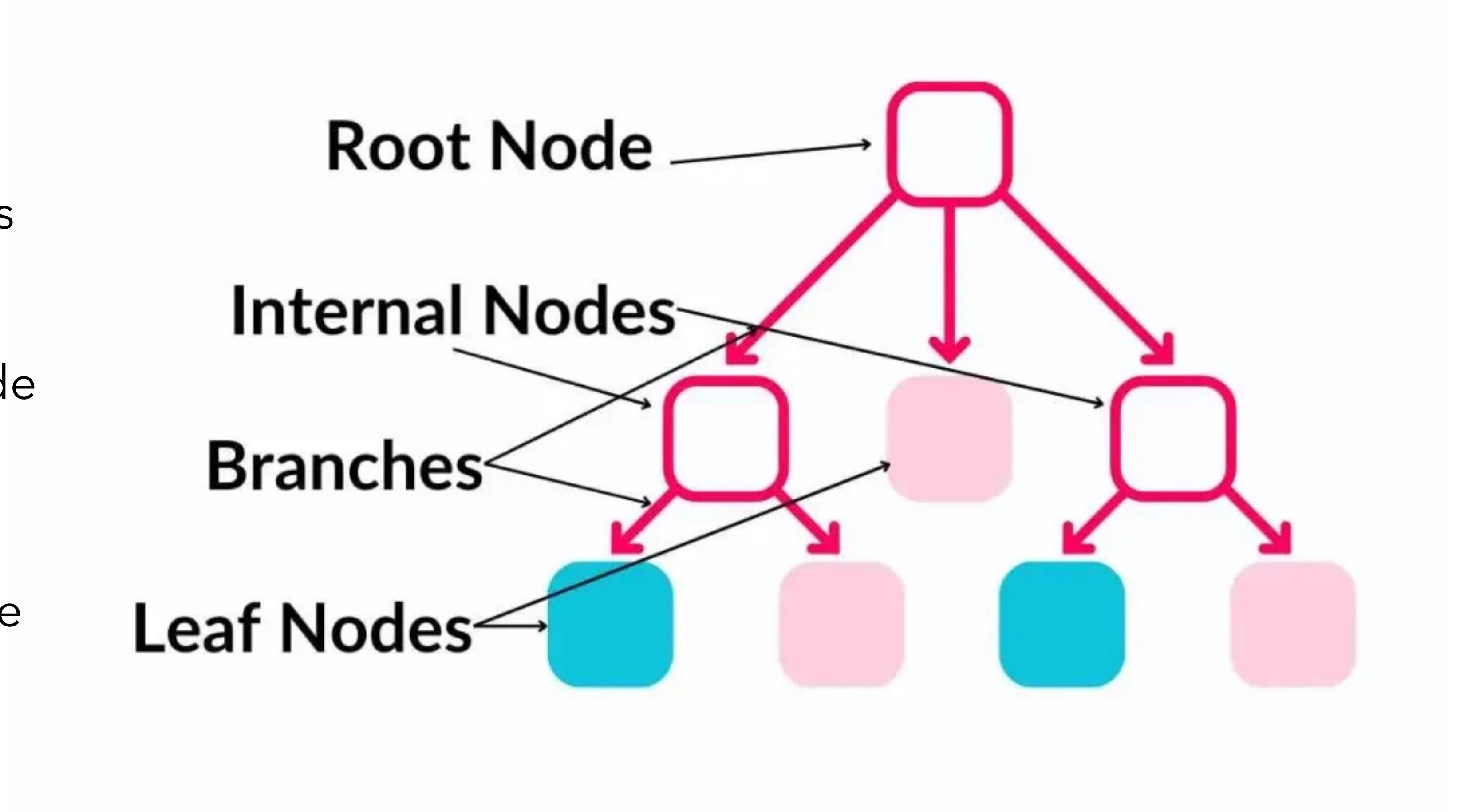
- Definition: A decision tree is a flowchart-like structure used for decision-making.
- Key Idea: Splits data into subsets based on feature values.
- Binary tree or multiple split tree



Source: https://medium.com/analytics-vidhya/decision-tree-end-to-end-implementation-adf1bc246254

Structure of a Decision Tree

- Root Node: The topmost decision node.
- Leaf Node: Terminal nodes representing decisions or outcomes.
- Branch: Path from one node to another.
- Splitting: Dividing data based on a feature.
- Pruning: Reducing tree size to avoid overfitting.



Source: https://spotintelligence.com/2024/05/22/decision-trees-in-ml/

How Decision Trees Work

- Classification Tree
  - Step 1: Identify a feature to split.
  - Step 2: Create branches based on feature values.
  - Step 3: Repeat until all data is classified or a stopping condition is met.
- Regression Tree
  - Step 1: Divide data to minimize variance within subsets.
  - Step 2: Predict outcome using the mean or median of leaf nodes.

Splitting Criteria

- For Classification:
  - Information Gain: Measures the unpredictability or disorder. Information gain is used to decide the optimal split by reducing entropy the most
  - Gini Impurity: Measures the frequency at which a randomly chosen element would be incorrectly classified. A lower Gini impurity indicates a purer node
- For Regression:
  - Variance Reduction: measure the reduction in variance for the dependent variable
  - Mean Squared Error
  - Mean Absolute Error

Stopping Criteria

- Maximum Depth:
  - Limits the depth of the tree.
- Minimum Samples for a Split:
  - Ensures a node must have a minimum number of samples before splitting.
- Minimum Samples for a Leaf:
  - Ensures each leaf node must have a minimum number of samples.
- Early Stopping:
  - Stops tree growth when the splits do not result in significant information gain.

Tree-Building Algorithms

- ID3: Uses Information Gain to decide splits.
- C4.5: Extension of ID3 that handles continuous data and missing values.
- CART (Classification and Regression Trees): Binary trees using Gini or MSE.

Algorithm	Splitting Criterion	Handles Numeric Data	Pruning	Tree Type
ID3	Information Gain	No	No	Multiple
C4.5	Gain Ratio	Yes	Yes	Multiple
CART	Gini Impurity / MSE	Yes	No	Binary

Advantage and Limitation

- Advantages
  - Simple to understand and interpret.
  - Handles both numerical and categorical data.
  - Requires minimal data preprocessing.
- Limitations
  - Prone to overfitting with complex trees.
  - Can be unstable with small data changes.
  - Bias toward dominant classes if data is imbalanced.

Overcome Limitation

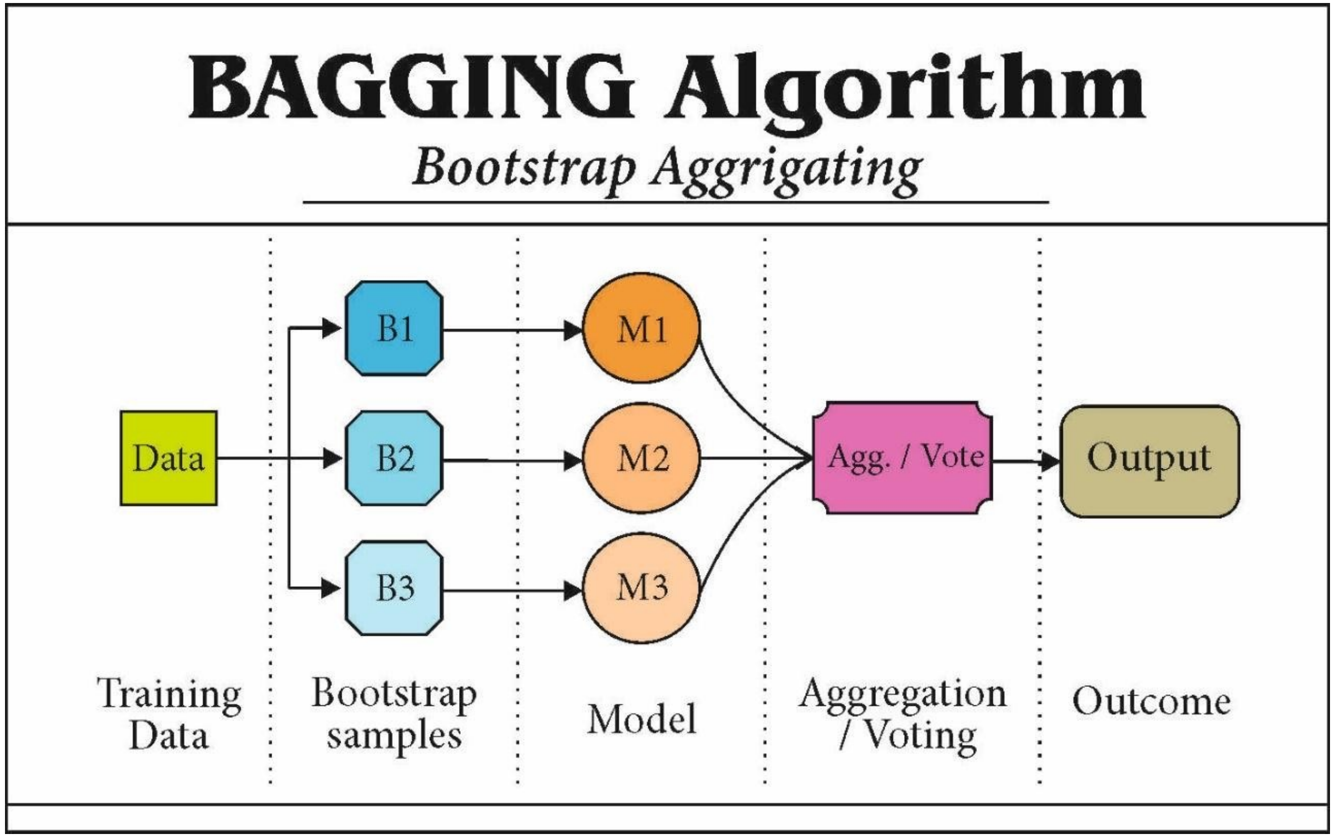
- Pruning Techniques
  - Pre-pruning (stopping early).
  - Post-pruning (removing branches after the tree is built).
- Ensemble Methods
  - Random Forest.
  - Gradient Boosted Trees.

Random Forest

What is bagging?

Definition: Bagging (Bootstrap Aggregating) is an ensemble method that combines the predictions of multiple models trained on different subsets of data.

Goal: Reduce variance and prevent overfitting by aggregating results.

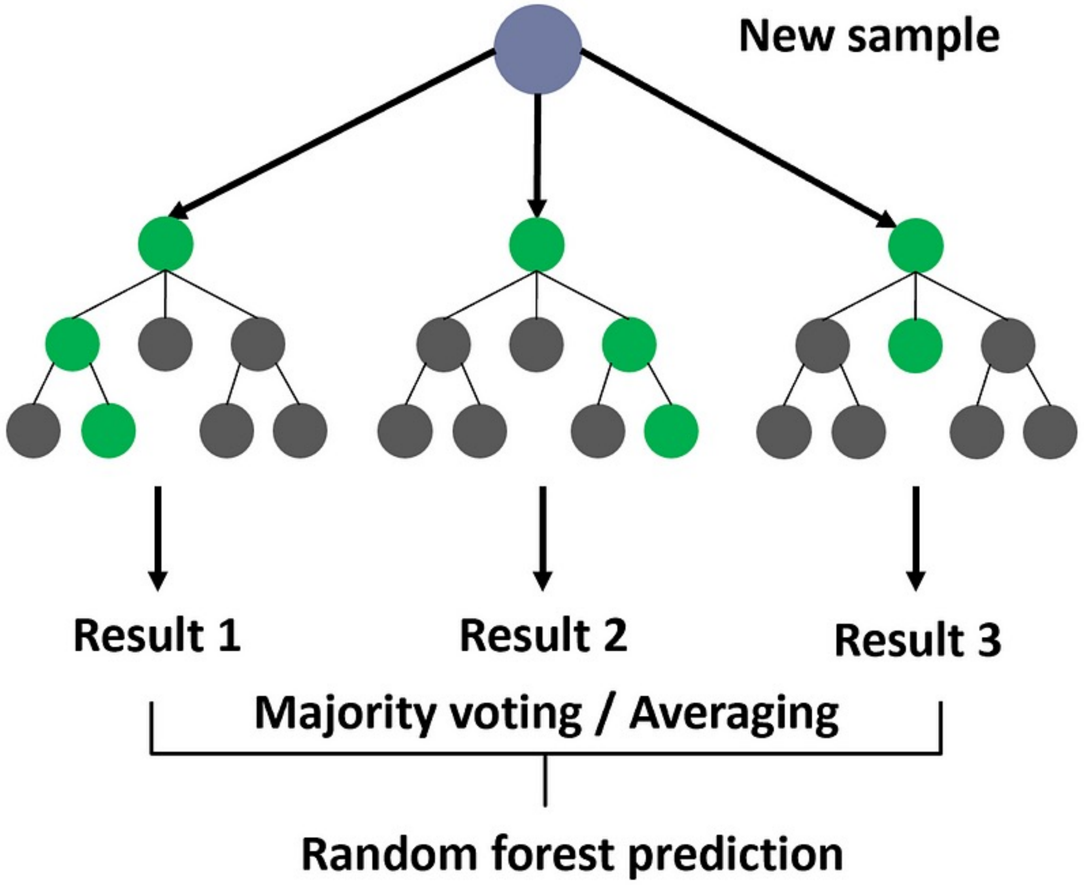


Source: https://www.linkedin.com/pulse/bagging-boosting-machine-learning-nagababu-molleti/

Random Forest

Definition: An ensemble of decision trees trained on different bootstrap samples with additional randomization in feature selection.

Key Idea: Combines multiple decision trees to improve predictive performance and reduce overfitting.



Graded Assignment

- Link: TBA
- Deadline: **Thursday, 11.12.2024**

References

- https://www.youtube.com/watch?v=_L39rN6gz7Y
- https://medium.com/analytics-vidhya/decision-tree-end-to-end-implementation-adf1bc246254
- https://spotintelligence.com/2024/05/22/decision-trees-in-ml/
- https://github.com/akash18tripathi/Decision-Trees-implementation-from-scratch/blob/main/Decision%20Tree%20Implementation.ipynb
- https://www.linkedin.com/pulse/bagging-boosting-machine-learning-nagababu-molleti/