

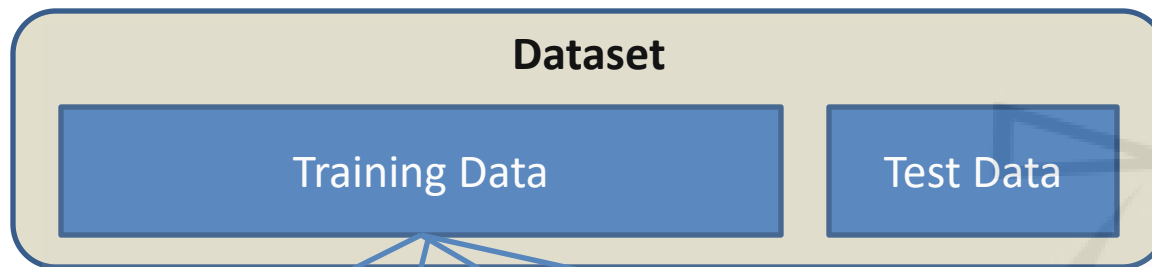


Random Forests

Mukesh Kumar

Random Forest Algorithm

- Bagging
- Random subspace method
- Training estimators
- Perform inference by aggregating predictions of estimators



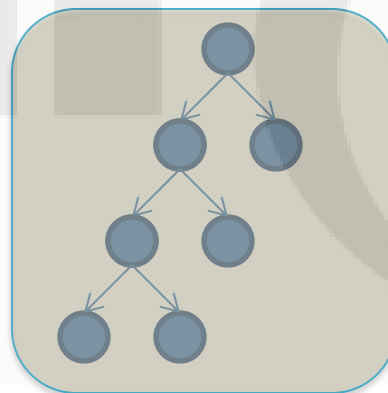
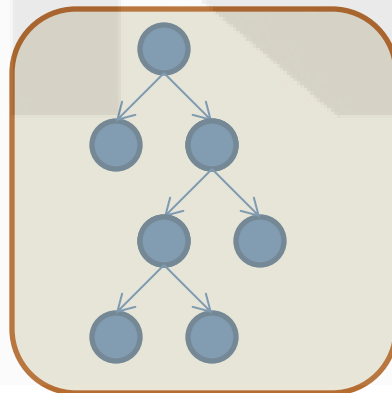
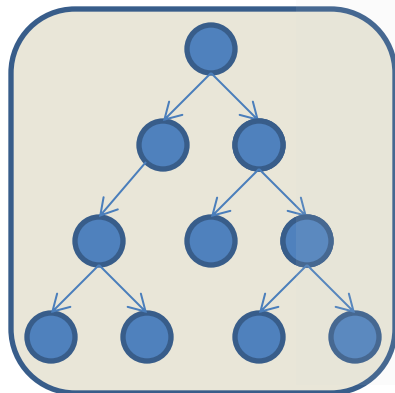
Data Subset 1

Data Subset 2

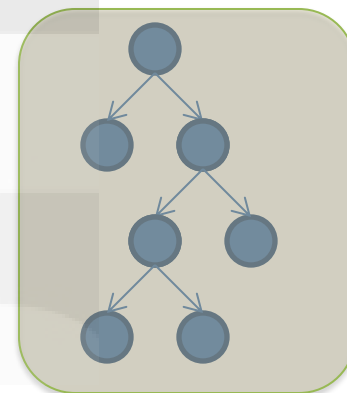
Data Subset3

.....

Data Subset n



.....



Result 1

Result 2

Result 3

Result n

Majority Voting/Averaging

How RF Works:

1. Bootstrapping
2. Feature Randomness
3. Building Trees
4. Voting/Averaging

Bootstrapping

- Random Forest creates multiple subsets of the training data through bootstrapping (sampling with replacement). Each subset is used to train a separate decision tree.

Feature Randomness:

Feature Bagging (Random Subspace Method):

- Feature bagging, also known as the random subspace method, refers to selecting a random subset of the features (or variables) for each tree to split on at each node.
- Rather than using all available features, each tree in the random forest only considers a random subset, which leads to **different decision splits** across trees.


Building Trees:

- Each tree is grown to the maximum depth without pruning, using the chosen subset of features and data.

Voting/Averaging:

- For **classification** tasks, each tree votes for a class, and the class with the majority of votes is chosen as the final prediction.
- For **regression** tasks, the predictions from all trees are averaged to produce the final output.

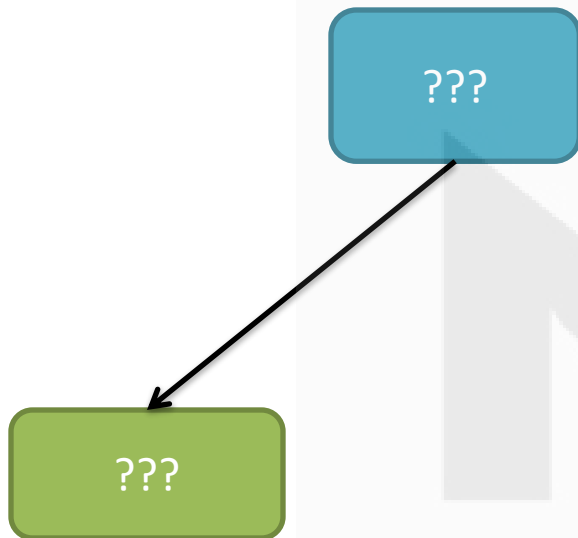
How RF builds tree



S.No	Loves Action Genre	Has Watched Top Gun	Age	Tom Cruise is Fav Actor
1	Yes	Yes	7	No
1	Yes	Yes	7	No
3	No	Yes	18	Yes
4	No	Yes	35	Yes
4	No	Yes	35	Yes
6	Yes	No	50	No
7	No	No	83	No

At every node it randomly picks “n” number of features for split where “n” is less than total number of features

How RF builds tree



S.No	Loves Action Gerne	Has Watched Top Gun	Age	Tom Cruise is Fav Actor
1	Yes	Yes	7	No
1	Yes	Yes	7	No
3	No	Yes	18	Yes
4	No	Yes	35	Yes
4	No	Yes	35	Yes
6	Yes	No	50	No
7	No	No	83	No

How RF builds tree

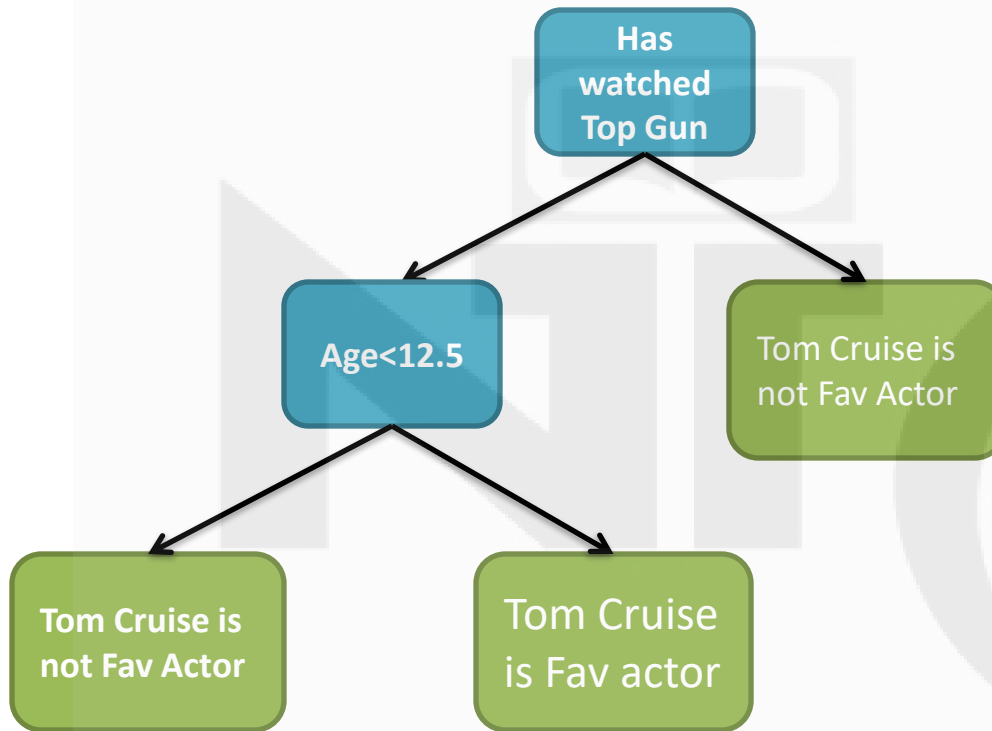
Has
Watched
Top Gun

???

S.No	Loves Action Gerne	Has Watched Top Gun	Age	Tom Cruise is Fav Actor
1	Yes	Yes	7	No
1	Yes	Yes	7	No
3	No	Yes	18	Yes
4	No	Yes	35	Yes
4	No	Yes	35	Yes
6	Yes	No	50	No
7	No	No	83	No

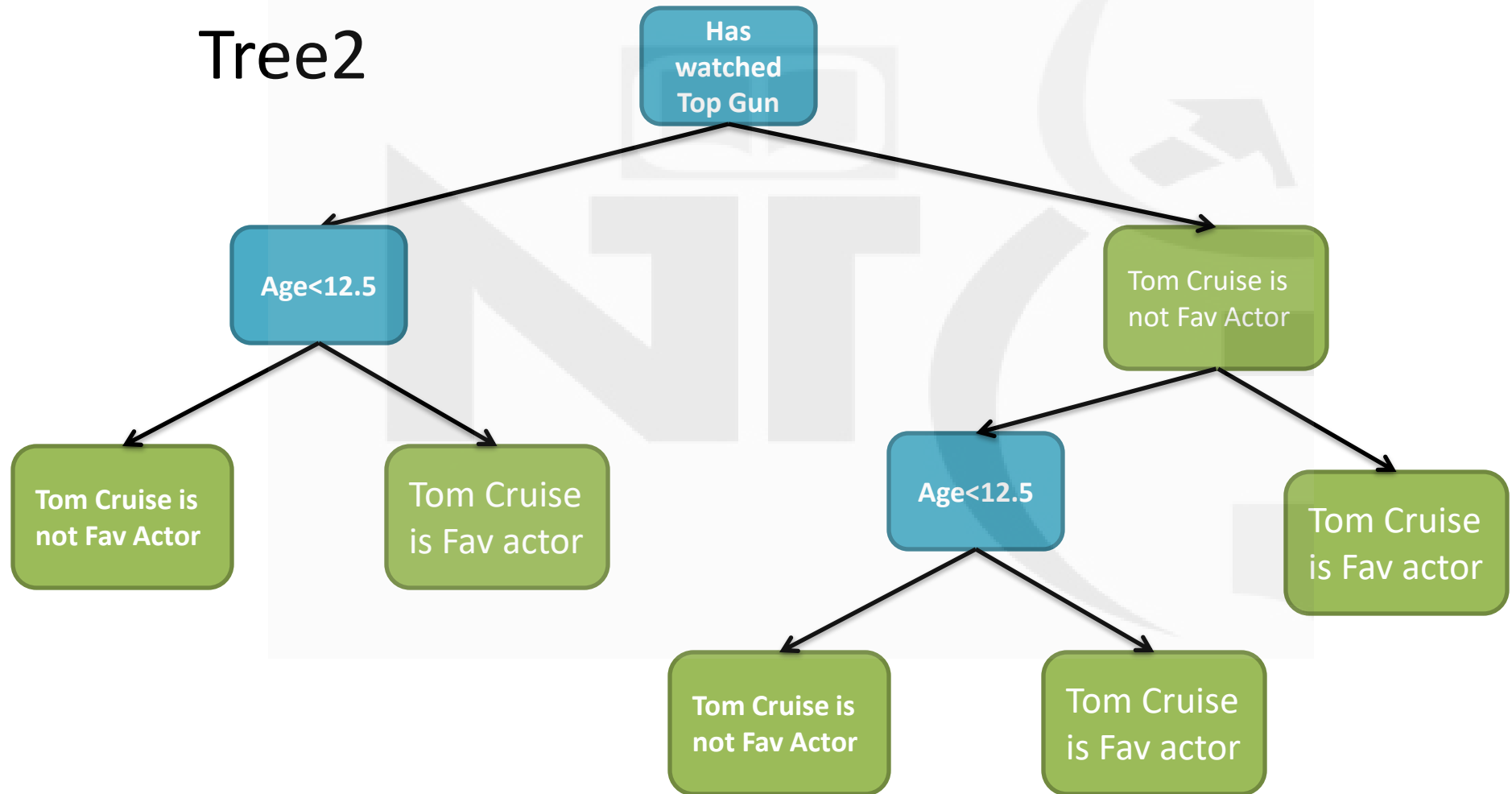
Since “Has Watched Top Gun” is taken , we do not consider that feature which choosing “n” features at second node

This is our Tree 1

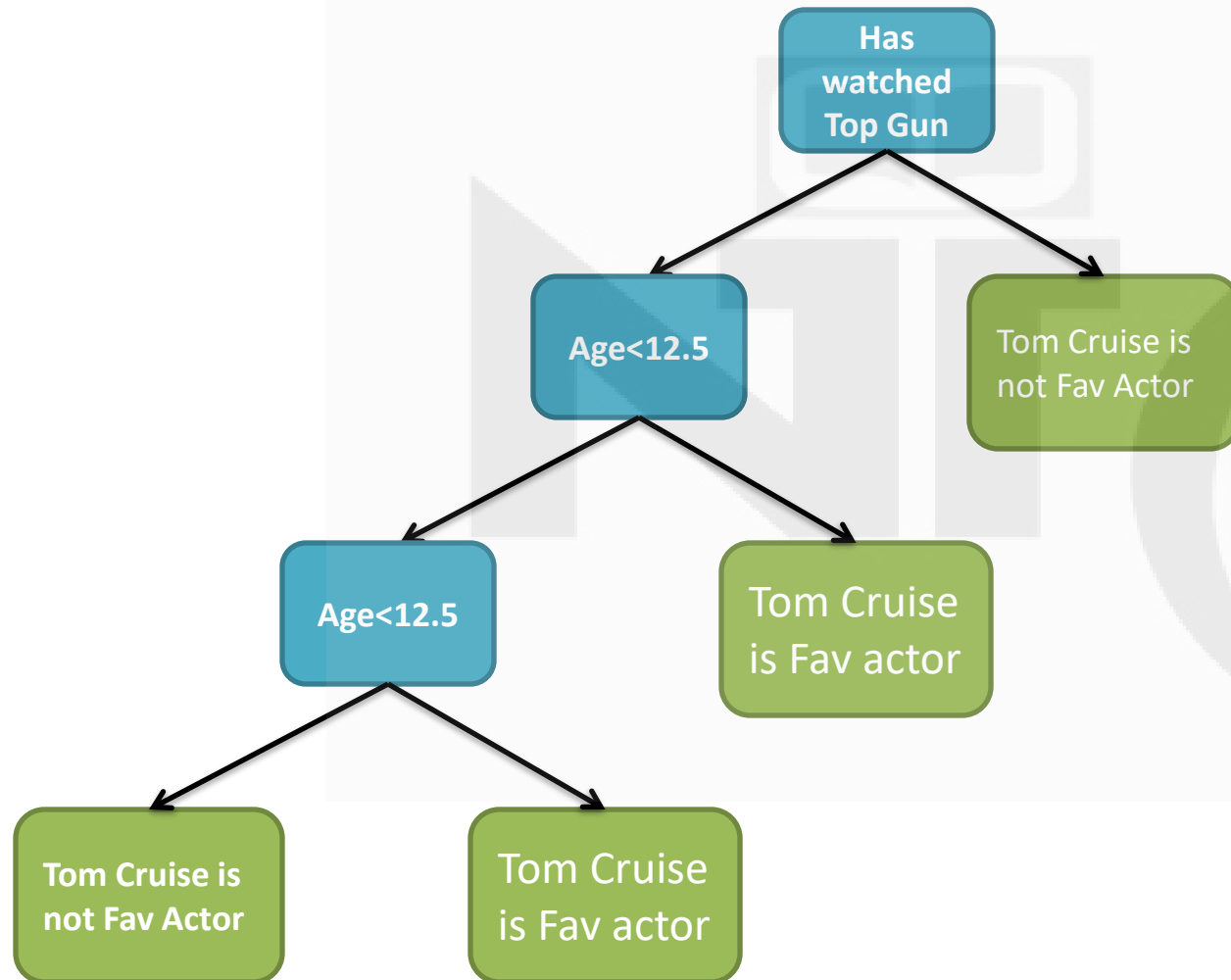


This process is repeated to build more trees

Tree2



This is our Tree 3



Feature Bagging (Random Subspace Method)

- If each training example has M features,
- we take a subset of them of size $m < M$ to train each estimator.
- None of the estimator sees the full training set, each estimator sees only m features of n training examples

Random subspace method

- **Node Splitting:** At each node of a decision tree, a random subset of features is selected, and the best split is determined only from this subset.
- **Repeat Process:** This process is repeated for every node in every tree, resulting in a variety of splits even for trees trained on similar or identical data.
- **Advantage:** Reduces correlation between trees by ensuring that different trees see different sets of features, even if trained on the same data.

Values of m, n

- **For Classification:** Usually, the number of features selected is the square root of the total number of features (\sqrt{m}).
- **For Regression:** Typically, a third of the total number of features is considered ($m/3$).

Benefits of Combining Both Techniques

- **Increased Model Diversity:** By combining bootstrapping with feature bagging, random forests create highly diverse models. This diversity is critical in reducing overfitting and improving generalization to new data.
- **Reduction of Overfitting:** Both techniques help prevent any single tree from becoming too complex and overfitting the training data.
- **Enhanced Robustness and Accuracy:** The ensemble of diverse trees tends to be more robust and accurate compared to individual decision trees.

Training estimators

- We create N_{tree} decision trees, or estimators, and train each one on a different set of m features and n training examples.
- **The trees are not pruned**, as they would be in the case of training a simple decision tree classifier.

Aggregating Outputs

- **Classification:** majority voting to decide on the predicted class.
- **Regression:** we will take the mean value of the predictions of all the estimators.

Why RF works well

- Trees are not correlated as they are built on a different subset with different features allowing each estimator to learn pattern in data from every feature's perspective

Disadvantages

- **Complexity and Interpretability:**
 - **Complex Model:** Random Forest consists of many decision trees, making the overall model more complex and harder to interpret than a single decision tree.
 - **Black Box Nature:** Understanding the contribution of each individual feature to the final prediction is difficult, which can be a problem in fields where model interpretability is crucial.
- **Computational Cost:**
 - **Training Time:** Training a Random Forest can be computationally intensive and time-consuming, especially with a large number of trees and high-dimensional data.
 - **Prediction Time:** Making predictions can also be slower compared to simpler models, as it requires aggregating predictions from all trees.
- **Memory Usage:**
 - **High Memory Consumption:** Storing a large number of trees can require significant memory, which might be a limitation when dealing with very large datasets.

Assignments

- Build a RF classifier for Titanic dataset
- For the Pima indians data set build boosting also and compare results with RF and decision tree
- Fine tune max depth for RF classifiers